COMPUTATIONAL MEASURES OF LINGUISTIC VARIATION: A STUDY OF ARABIC VARIETIES

BY

MAHMOUD ABEDEL KADER ABUNASSER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Elabbas Benmamoun, Chair
Professor Mark Hasegawa-Johnson, Co-Chair
Professor Ryan Shosted
Professor Eiman Mustafawi

ABSTRACT

This thesis introduces and discusses a new methodology for measuring the variation between linguistic varieties. I compare five Arabic varieties – Modern Standard Arabic MSA, Gulf Arabic GA, Levantine Arabic LA, Egyptian Arabic EA, and Moroccan Arabic MA – considering both lexical and pronunciation variation. I introduce the idea of measuring the amount of linguistic variation asymmetrically; the amount of linguistics variation between a speaker of variety A and a hearer of variety B is not necessarily equal to the amount of linguistic variation between a speaker of variety B and a hearer of variety A. I propose a new mathematically based computational representation of sound that enables the incorporation of phonetic features and articulatory gestures in measuring the amount of pronunciation variation. I also implement an optimization technique to assign weights and parameters to the phonetic features and articulatory gestures for the proposed representation of sound. The developed methodology, tools and techniques lead to a better understanding of the structure of language and have implications for both theoretical linguistics and applied work in natural language processing NLP, it both provides a computational technique to assess the plausibility of defining the components of sound and opens a new venue to the possibility of utilizing a representation of sound that is phonetically motivated and computationally applicable to NLP problems. This research could potentially yield insights into the issues of mutual intelligibility between Arabic varieties and dialect identification.

Measuring lexical and pronunciation variation is based on native speaker elicitations of the Swadesh list for the local varieties of Arabic; MSA is represented by data from dictionaries. The data collection procedure allows the participants to provide more than one translation. I also provide a context sentence for all lexical items to rule out cases of ambiguity. The amount of

lexical variation is measured at two levels of representation: the word level and the phonemic level. At the word level, the amount of linguistic variation is based on whether the words share a linguistic origin. The phonemic level, using IPA transcription of words, looks at more details in measuring the lexical variation. The amount of pronunciation variation is measured at three levels. The first and most abstract level is the phonemic level. The second incorporates the mathematical representation of sound; which encodes phonetic features and articulatory gestures. The third allows the vowels to be represented non-categorically based on the values of the first and second formant frequencies, MSA is not included at this level.

The results of the measures of linguistic variation developed in this study confirm two observations about the communication between speakers of the Arabic varieties and provide an answer for the frequently asked question about the closeness of the Arabic varieties to each other. The first observation is that MA seems to be relatively distant from the other local varieties (GA, EA, and LA) than those varieties are from each other, which relates to the geographical distances between those varieties. The second observation is the asymmetric pattern of intelligibility in the communication of EA speakers with the members of the other local varieties; GA, LA, and MA speakers seem to understand EA speakers better than the EA speakers understand them. This asymmetric pattern of intelligibility is reflected by the variation metrics developed in this research. As for the closeness of the local varieties to MSA, GA and – to some extent – LA seem to be the closest, followed EA, and MA is the farthest. In addition, EA seems to be closer to MA than both LA and GA. Moreover, EA speakers are closer to LA hearers than GA hearers. On the other hand, GA speakers are closer to LA hearers than EA hearers. Finally, the last measure, that of pronunciation variation, situates LA speakers closer GA hearers than EA hearers.

ACKNOWLEDGEMENTS

Thanks are due to God for his endless blessings upon me. My profound gratitude is to those who have supported and helped me throughout my education. Among the many people who have helped me, two have provided help during my graduate studies beyond what is expected from them: my advisor Elabbas Benmamoun and my friend and colleague Tim Mahrt. Without the help I have received from them, I would have not accomplished this. I would like to thank them for their generous help, patience, support, and guidance. I am forever grateful to Mark Hasegawa-Johnson for his support and for his flexibility which allowed me to craft what I consider the most important piece in this work: the linguistically motivated, mathematically grounded, and computationally effective representation of sound. No doubt, Ryan Shosted provided great insight in the area of phonetics. I was able to better flesh out a non-categorical representation for vowels with his valuable feedback and comments. I would like to extend my gratitude to Eiman Mustafawi for the well-thought and accurate questions and comments about the representation of Arabic varieties and the transcription procedure; sharing her expertise was a great help for my research. In addition to my respected committee members, I have discussed ideas related to the research of this dissertation with other professors. I would like to thank Chilin Shih, José Hualde, Jennifer Cole, and Mona Diab for their valuable feedback.

Most importantly, my deep appreciation goes to my wonderful parents who raised me in my childhood and are still caring, loving, and supportive. I would have certainly not be where I am now without their sincere devotion. And my beloved wife Halema is the one that shared with me every single moment in this journey. Halema has made my success throughout graduate school possible with her love, support, and determination. My wonderful children, Ahmed, Mira,

and Yamen were the biggest motivation during my studies. I thank my sister Eman for encouraging me the most to pursue a PhD degree. My brother Ahmad has always been supportive. I thank him for all the useful discussions and for always standing by my side in this long journey. I also would like to thank my brother Mohammad who have helped me better understand optimization techniques used in Engineering. My sister Enas has been supportive and encouraging. I thank her for everything she did for me and for my family.

I am in debt to the participants who gave their time and patience to complete the required recordings for this research. I am also grateful to my friends Abdelaadim Bidaoui and Iftikhar Haider for their support and help. I will miss our conversations and the tea we had together on an almost daily basis. I am also grateful to my friend Moad Hajjam and my niece Raneem Saadah for making drawings for one of one of the pilot studies I used to prepare for this research. I also would like to thank Daniel Ross for his willingness to review the manuscript. His help enabled a better flow of the ideas in addition to correcting spelling and grammatical mistakes. I owe Tim Cunningham my gratitude for the excellent IT support he has provided throughout my stay at the U of I.

I gratefully acknowledge Qatar National Research Fund (QNRF) grant NPRP-09-410-1-069 (M. Hasegawa-Johnson, PI) and National Science Foundation (NSF) grant BCS-0826672 (E. Benmamoun, PI) for partially funding this research.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION AND OVERVIEW

Measures of linguistic variation, also called linguistic distance, is one of the prominent topics in the growing field of dialectometry, which is concerned with quantifying linguistic differences and similarities and, often, relates it to geographical distances between the areas where the relevant languages/varieties are spoken (Nerbonne and Kretzschmar 2003). In this thesis, I report on a set of computational measures of linguistic variation that quantifies the lexical and pronunciation variation between five Arabic varieties: Modern Standard Arabic MSA, Gulf Arabic GA, Levantine Arabic LA, Egyptian Arabic EA and Moroccan Arabic MA. The drive to computationally study linguistic variation is partly due to the extensive typological literature and the increasing number of corpora from different languages, which makes this type of research possible. Dialectometry has the potential to enrich the debates in a variety of fields such as theoretical linguistics and its focus on microvariation and its extents and limits as well as the related issues it raises about the cognitive aspects of language, in addition to anthropology, sociology and history, among many others.

This research provides empirical evidence regarding the amount of linguistic variation between the Arabic varieties under consideration. Hence, it provides an answer for the frequently asked question about the closeness of the local varieties to MSA. Moreover, it provides empirical evidence based on computational techniques for two observations about the linguistic communication between speakers of the local Arabic varieties. The first observation is that MA is more distant to the other local varieties of Arabic considered in this study than the other varieties among themselves; Geographically, MA is also more distant. The second observation is

that, in most cases, Egyptian speakers are understood by other varieties better than they understand them. It is important to note that this observation may be due to factors related to exposure to Egyptian media, which is popular in many other countries. Also, it may be due to factors related to the linguistic competence of the speakers on both sides. Of course, the former might have effect on the latter, for example, exposure might result with lexical items to be borrowed from one variety to the other, which become part of the linguistic competence of the speakers of both varieties. This research provides evidence about the amount of linguistic variation between the varieties as they are currently spoken. The questions about the reasons that might affect the amount of variation, such as exposure, are outside the scope of this research.

The term linguistic distance has been extensively used in the field of dialectometry to express the amount of linguistic variation between varieties. However, this term is problematic as 'distance' implies a single measure calculated between two objects. As shown by the use of the term mutual intelligibility, the measure of intelligibility is inherently asymmetric, meaning that speakers of some variety (A) may understand speakers of another variety (B) better than speakers of variety (B) understand speakers of variety (A). In this thesis, I develop variation metrics that are asymmetric. Instead of the term linguistic distance, I am using the terms measure of linguistic variation and linguistic variation metric; they are used interchangeably in this thesis.

Séguy was among the first researchers in the field of dialectometry. In his 1973 study, he used a linguistic Atlas that contained variables from five linguistic subsystems or components that represent the languages under consideration. The linguistic subsystems were lexical (represented by 170 variables), pronunciation (67), phonetic/phonological (75), morphological (45), and syntactic (68). For each subsystem or component, Séguy calculated the percentage of disagreements between each neighboring pair of sites for each variable in the five subsystems.

Then the linguistic distance is calculated as the average of the distances between the five subsystems (Heeringa 2004).

In this research, I study each linguistic subsystem independently when measuring linguistic variation, which in our view is the most efficient, informative and feasible way to measure linguistic variation. For the present purposes, the scope of the investigation is limited to two linguistic subsystems: lexical and pronunciation. Other subsystems, such as morphology, morphosyntax and semantics, are to be studied in the future. It is important to explore each linguistic subsystem independently because the amount of linguistic variation in each subsystem might have different implications. For example, from a Natural Language Processing (NLP) point of view, greater variation in the lexical subsystem indicates more differences in a dictionary to be used in an automatic translation system. A smaller variation in pronunciation might imply that an automatic speech recognition system trained on one dialect is usable, to some extent, for the other dialect. Similarly, morphosyntactic and morphological distance should reflect the amount of adaptations or changes required to make a morphological analyzer or stemmer usable for the other variety.

The question of measuring linguistic variation has been approached from different perspectives. Some studies have looked at the distance between languages in an effort to reconstruct the languages family trees (Gray and Jordan 2000; Gray and Atkinson 2003; Serva and Petroni 2008, among others). Others have looked at the distance between closely related languages, or dialects of the same language, in an attempt to identify the subgrouping of those languages or dialects (Elsie 1986; Ebobisse 1989; Babitch and Lebrun 1989; Kessler 1995; Heeringa 2004; Valls et al. 2011, among others). Yet another stream of research has employed measures of linguistic variation in computational tasks such as the automatic identification of

cognate words (Kondrak and Sherif 2006; Kondrak 2009). Gooskens (2007) tested the correlation between different measures of linguistic variation and mutual intelligibility between Scandinavian languages to show that the amount of phonetic variation can predict the degree of mutual intelligibility better than the amount of lexical variation. Within the area of Arabic linguistics, the most relevant area of research has been concerned with the problem of dialect identification (Biadsy et al. 2009; Zaidan and Callison-Burch 2012; Elfardy and Diab 2013).

The motivation for this study is to enhance our understanding of linguistic variation and thereby enhance our understanding of human language as a whole. This is based on the idea that quantifying the amount of variation between two entities enforces a better understanding of the nature of the entities under consideration.

The goals of this study are both conceptual and empirical. Conceptually, I develop a representation of sound that captures phonetic similarity in a mathematically simple and computationally feasible way. This representation of sound is based on phonetic features and articulatory gestures; it is an attempt to computationally represent the sound based on its basic components. It is also equipped with the ability to represent sound categorically and non-categorically, this representation of sound is referred to as the mathematical representation of sound. The second conceptual goal is to provide a non-subjective way to assign weights to phonetic features. The first two conceptual goals are crucial to answer the question of how to computationally measure pronunciation variation. Which in turn, leads to models and techniques that could potentially help solve problems related to similarity in pronunciation raised in various NLP tasks. The third conceptual goal is to introduce the idea of measuring linguistic variation asymmetrically. This is important to solve the puzzle of asymmetric mutual intelligibility. Empirically, I develop a set of techniques to measure the amount of lexical and pronunciation

variation between closely related (and possibly mutually intelligible) languages. I incorporate data from four local varieties of Arabic and from MSA to measure the lexical and pronunciation variation among them. I provide a new approach to computationally handle pronunciation variation based on a mathematical representation of sound. I also consider which features should be included in the representation of a given sound, and the salience of each of these features. I measure the amount of linguistic variation between all pairs of Arabic varieties included in this study, which answers the frequently asked question about the closeness – here, in terms of lexical and pronunciation variation – of the local varieties to MSA. It is important to keep in mind that I focus on the amount of variation between MSA speakers and hearers from the local varieties, which reflects the ability of the members of local varieties to comprehend MSA. The other direction of communication is not highlighted in the discussion because it relates to the ability of MSA native speakers to comprehend the local varieties; the existence of MSA native speakers is questionable and if exists their ability to comprehend the local varieties would not be of a high cultural and social importance.

The primary guideline in making decisions related to the data analysis and the design of the data collection procedure is to mirror the degree of mutual intelligibility between two speakers when they are first encountered or after a limited exposure. It is important to note that the amount of linguistic variation that we are measuring is not the only factor that affects the degree of mutual intelligibility. Exposure is another factor or perhaps one of the most important factors that facilitates mutual intelligibility. Speakers from different dialects maybe exposed to each other and may develop some familiarity with each other's dialects. Even if someone is not exposed to some dialect he/she might be exposed to another dialect that has some features that exist in the first dialect. For example, a speaker of the dialect spoken in Cairo, Egypt, does not

have gender agreement in verbs for third person plural verb subjects in his/her EA grammatical system. However, since he/she might be exposed to Standard Arabic, which has that feature, we do not expect to see significant intelligibility problems with respect to third person feminine plural agreement with speakers of some GA dialects which have that grammatical feature.

The four local varieties are represented by elicitations of the words of the Swadesh list from two native speakers born and raised in a major city where the variety is spoken. MSA is represented by translations of the words of the Swadesh list from two dictionaries of MSA (see chapter 2). The lexical subsystem is investigated at two levels of representation, the word level, and the phonemic level. At the word level, the amount of linguistic variation is based on whether the words have originated from the same linguistic origin. The phonemic level considers more details by measuring the lexical variation based on the similarity of the IPA transcription of words of the Swadesh list. The pronunciation subsystem is investigated at three levels. The first and most abstract level is the phonemic level. At this level, we measure the amount of pronunciation variation based on the similarity of the IPA transcription of cognate words in the Swadesh list. The second level incorporates the mathematical representation of sound which takes into account the phonetic features and articulatory gestures in measuring pronunciation variation. The third level allows the vowels to be represented non-categorically based on the values of the first and second formant frequencies. MSA is not included in the third level due to the lack of acoustic data.

All measures that took into account MSA have situated MA as the farthest to MSA. The lexical measure at the word level resulted with LA as the closest to MSA, followed by GA then EA. The remaining three measures have situated GA as the closest to MSA. Two of them had LA in the second place. As for the variation between the local varieties, the closest to MA is EA

followed by GA and LA. The variation metrics did not provide a significant distinction between the closeness of GA to MA and the closeness of LA to MA. All variation metrics showed that GA speakers are closer to LA hearers than EA hearers. The lexical measure at the phonemic level and the first two pronunciation measures showed that EA speakers are closer to LA hearers than GA hearers. On the other hand, the third measure of pronunciation variation showed that LA speakers are closer to GA hearers than EA hearers. See Chapter 6 for more discussion about the closeness of the Arabic varieties to each other.

For many studies in dialectometry, the focus is categorizing different dialects into subgroups (Elsie 1986; Babitch and Lebrun 1989; Ebobisse 1989). One shortcoming in this approach is that the focus often drifts to defining dialect boundaries, which is not the focus of the current research. Séguy (1973) introduced the idea of providing a distance matrix that replaced the method of counting the number of isoglosses between dialect sites and ruled out the problem of dialect subgrouping. In this project, I follow Séguy (1973) by providing results in a distance matrix as opposed to providing the results on a map. The distances reported by each metric are best interpreted relative to other results from the same metric, reported in the same table.

The rest of this thesis is organized as follows. Chapter 2 discusses the data, the data collection procedure and the preparation of the data for the use of the measures of linguistic variation. Chapter 3 reports on the measure of lexical variation at the word level. Chapter 4 describes measures of lexical and pronunciation variation at the phonemic level. Chapter 5 discusses the mathematical representation of sound and respective methodology used in measuring pronunciation variation. The conclusions, limitations, and implications of this research and future directions are discussed in Chapter 6.

CHAPTER 2

PARTICIPANTS, DATA SOURCES, AND DATA COLLECTION PROCEDURE

This chapter covers all the steps required to prepare the data for measuring the lexical and pronunciation variation between the varieties of Arabic. The first section discusses the data sources used to elicit the words of the Swadesh list. Each local variety is represented by two male native speakers born and raised in a major city where the variety is spoken. MSA is represented by two modern dictionaries of Arabic. The second section reviews the Swadesh list and discusses its usability for the Arabic varieties where we found that some adaptations are required. For example, some meanings are clarified or restricted by context sentence. The third section touches the issue of allowing the participants to provide more than one translation for the items in the Swadesh list. The data collection procedure and tools developed to facilitate the data collection are discussed in the fourth section. The data segmentation and transcription are discussed in sections five and six respectively. Section seven reports on the algorithm I developed to predict landmarks at which the values of the formant frequencies are sampled. The last section discusses a non-categorical representation for vowels based on the values of the first and second formant frequencies. The remaining chapters in this thesis discuss the procedures and methods to measure the lexical and pronunciation variation between the varieties of Arabic based on the data sets prepared according to the methods described in this chapter.

2.1 Participants and MSA Data Sources

Each spoken variety is represented by two male native speakers between the ages of 21 and 32. All participants were required to have been born and raised in a major city where that variety is spoken; their parents must also speak the same dialect. For this study, we only consider male speakers in order to eliminate any possible effect of gender in the data. I tried as much as possible to have all participants of similar socio-economic status from the middle class. More information about the participants is provided in Table 2.1.

Table 2.1: Summary of the participants

Dialect	ID	City	Social status	Year of Birth
EA	EA01	Cairo, Egypt	Middle, upper	1983
EA	EA02	Cairo, Egypt	Middle	1982
GA	GA01	Dharan, Saudi Arabia	Middle	1982
GA	GA02	Manamah, Bahrain	Middle	1984
LA	LA01	Salt, Jordan	Middle	1984
LA	LA02	Tripoli, Lebanon	Middle, upper	1992
MA	MA01	Meknes, Morocco	Middle	1982
MA	MA02	Rabat, Morocco	Middle	1982

MSA is represented by two modern dictionaries, namely *Almawrid* (Baʿalbaki and Baʿalbaki 1999) and *Elias Modern Dictionary* (Elias and Elias 1983). Because MSA is a standardized language, the lexical items from these dictionaries are considered an accurate representation of the language. One complication was that the dictionaries listed some dialectal forms such as *ʔɛːʃ* ʿwhat' from the Levantine dialect *haraʃ* ʿrub' from the Egyptian dialect. Therefore, these words were removed from the data set after consulting other modern and classical dictionaries of Standard Arabic (*muxtaar ʔassiħaaħ*, *lisaan ʔalʕarab*, and *ʔassiħaaħ fi ʔalluʁa*). Also, because the lexical items in the Swadesh list sometimes had multiple possible

translations, I selected only the translations that matched the context assigned to the items (see Section 2.2).

2.2 Swadesh list

The Swadesh list is widely used in linguistic research. The list consists of 207 lexical items that contain different parts of speech including pronouns, nouns, adjectives, verbs, prepositions and others. The Swadesh list is provided in Appendix A including the translations from the Arabic varieties and the original English version. Some adaptations are introduced to the list to make it usable for Arabic varieties and to eliminate, as much as possible, the effect of the other linguistic subsystems on the lexical and pronunciation variation. These adaptations are achieved by introducing the word in a context sentence. To be consistent, all words are given context sentences even if the context is not necessary. The first and most frequent adaptation is to select a single verbal form with the same tense and person, gender, and number agreement for all verbs. This is necessary to ensure that the participants are not providing different inflections or tenses for the verbs. All verbs were elicited in the past tense with third-person masculine singular agreement, which has no prefixal or suffixal inflections and is the conventional form listed in Arabic dictionaries. This eliminated as much as possible the effect of the morphosyntactic subsystem. Likewise, the masculine form was selected for the two instances of the pronoun *you* (singular and plural) for consistency. In addition, nouns were elicited in an indefinite (unmarked) form and cliticized pronouns were removed. Moreover, word final vowels are not included for verbs, nouns, adjectives, or quantifiers, for which the vowel in most case indicate grammatical inflection rather than lexical information. Other lexical categories such as pronouns, demonstratives, question words, negation particles, prepositions, and conjunctions

were elicited with the word final vowel maintained because the word final vowel does represent lexical information for these classes.

The second adaptation is to provide a context sentence that disambiguate the meaning of some words in the Swadesh list. Although a context sentence is provided for each item, the disambiguation is necessary in three main situations for certain words. The first situation is syntactic where the translation of the item depends on the syntactic position of the word in the sentence. The negation particle *not*, appearing as item number 16 of the Swadesh list, can be translated in GA as *maa* to negate a verb and as *muu* or *miʃ* to negate a participial or an adjective.[1] In such cases, it is important to provide all participants with a single context to ensure consistency. The second situation is based on lexical semantic factors where the translation of the word is highly dependent on the context. For example, adjectives can be translated differently when they modify different nouns. The adjective *wide* in English can be used in *wide road* and *wide pants.* The translation of *wide pants* in EA is *bantʿaluun waaseʕ*. While *wide road* could be translated as both *ʃareʕ waaseʕ* and *ʃareʕ ʕariidʿ*. To avoid the situation where the participants are translating the adjectives with different contexts in mind, an explicit context is provided. Likewise, a preposition may have more than one spatial or temporal meaning. For example, *I am at home* and *I am at the door* denote different spatial meanings; the first denotes that the entity referred by the subject of the preposition is inside the home, whereas the latter means that the entity referred by the subject of the preposition is close to the door. The third situation is when a word has more than one meaning. For such cases, the context sentence ensures that all participants are translating the same word sense and avoids ambiguity. Examples from the

[1] In some varieties of Gulf Arabic, one can find *miʃ* in addition to the traditional Gulf negatives *maa*, *muu* and *mub*. *miʃ* is most likely a borrowing due to contact with varieties of Arabic, such as Egyptian, where *miʃ* is the typical negative particle.

Swadesh list include the verb *lie* which means to rest on a flat surface or to speak falsely. The noun *bark* might refer to the covering of a tree or to the sound of a dog. Similarly, *fat* might refer to the white residue in meat or to an overweight person. I selected the word sense that goes in line with the data provided by earlier studies that have used the Swadesh list.

In addition to the adaptations mentioned above, it was necessary to introduce some adaptations to a set of items in the Swadesh list. These adaptations are based on the researcher's experience in working with participants from the spoken varieties. The translation of item number 40, that corresponds to *wife*, was elicited in the construct state form (the so-called Idhaafa). Some participants provided an exclusively formal (MSA) translation for the word when it is not in Idhaafa construction, so the word in Idhaafa is considered more natural. It was also problematic to elicit a translation for item number 46, corresponding to *bird*. The size of the bird plays distinctive role in the translation of the word, so the context sentence specified the size of the bird. The class of demonstratives (items 7-10) was given as a topic of the sentence, and the participants were asked to utter it while pointing to the intended object. The coordination item *and* (number 204 in the Swadesh list) was produced by the participants in many different ways, including different ending vowels. In many cases, the same participant provided more than one form. Examples of the pronunciations include: *ʔu*, *wa*, *wu*, *wi*, and *u:*. To resolve the issue of extensive optionality, I asked the participants to add an epenthetic glottal stop and pause before and after uttering the item. The context sentence for this item is: *Ali ___ Saleh are friends*. The direction was given as follows: say the first name then pause. Start the coordination element by starting with an *ʔ* sound if needed, then pause again after uttering the coordinating element. Then say the second name. The pauses enforced the pronunciation of the item as a word rather than a prefix. This strategy worked very well to eliminate the variations caused by different optional

pronunciations of the same item between the native speakers of the spoken varieties. However this caused a problem comparing these different pronunciations to MSA because the selected pronunciation of the same item is not possible in MSA, for which the standard form *wa* was used.

2.3 Allowing Multiple Translations

The words of the Swadesh list were elicited in two passes. In the first pass, participants were asked to translate the words from English to their variety of Arabic. In the second pass, participants were given the words in the other varieties, in addition to the English form. The researcher discussed with the participants the possibility of using one of those words or words with similar linguistic origin in their variety used in the same context. The purpose of the first round is to find the most natural translations that the participants would provide without seeing what other participants have provided. The purpose of the second round is to find any possible optionality where a cognate of a word in one of the other varieties is available in the participant's variety with the same meaning. It is worth mentioning that some participants have said, in some cases, that the words they saw in the second pass have reminded them with a more natural translation of the English word[2]. All words are provided in Appendix A. The words that the participants provided in the first round are tagged with ENG, abbreviation of ENGlish. The words elicited in the second round are tagged with VAR, abbreviation of other VARieties. This requirement complicates the data collection procedure because each participant must be aware of all the words provided by the other participants. If a participant adds a new set of words, then all other participants have to be consulted about the newly added words. To simplify the process I

[2] An example is the translation of the word 'guts' (Swadesh item 86) provided by the speaker GA01. After seeing the translation provided by the other speakers, he said that *masʕArIn* is a better translation that the word *ʔamʕAʔ*.

anticipated what the participants would potentially provide by informally consulting native speakers from the varieties under consideration and by consulting an online resource.[3] If a participant introduced a word that did not exist in the list of possible words, the new word was added to a list that I used to reinterview all previously recorded participants.

2.4 Data collection procedure and tools

The data collection procedure is designed to allow the participants to provide translations of the English word along with a context sentence before they are shown the words in the Arabic varieties. It is also possible to add to the list of words in the Arabic varieties as I elicit data from the participants. For each item in the Swadesh list, the participant was first given the word in English along with the context sentence. Then the researcher discussed possible translations in his variety. The participants were always reminded that they must only provide words that they produce in informal settings such as when talking to siblings and close friends from the same city. When the participant is ready, he was asked to repeat each translation that fits the context sentence three times. After that, the participant was given a set of possible translations in the Arabic varieties according to the preliminary data I collected about the other varieties and according to what the previous participants have provided. If a cognate of any of the words exists in his variety and would be used in daily life for the given context then the participant was asked to repeat each of these new possible words three times. In many cases, the participants would say that they understand the word and they might have heard it spoken by speakers from their city, but they do not feel that they would say it themselves. In such cases, the word was not considered. In the event the participant added a new word that did not exist in the precompiled

[3] http://en.wiktionary.org/wiki/Appendix:Arabic_Swadesh_list

list of possible words in the Arabic varieties then the new word is added to the list so the following participants are asked about it. Here it is important to keep track of any added items. After I recorded the data from all participants, I recompiled the list with including the newly introduced lexical items and ask each participant about the items other participants have introduced after their first session.

The data collection was facilitated by *BrowseHTMList*, an application I developed to help manage the process of data collection. For phonetic analysis and manipulation of audio files, the *Praat* software was used (Boersma and Weenink 2012). This tool is an open source program used by the linguistic research community. The software was also used for the analysis of the recorded stimuli as discussed later in this chapter. The recordings took place in a sound-proof booth at the Phonetics and Phonology Lab at the University of Illinois at Urbana-Champaign using a Marantz digital recorder (Marantz PMD570) and an AKG c520 head-worn condenser microphone. The recordings were sampled at 48.0 kHz.

2.4.1 BrowseHTMList application

BrowseHTMList is an application developed by the researcher using MS Visual C++ 2005. Its main function is to load a list of HTML pages, each page is associated with an ID. The application allows a user to browse through the HTML pages in the order they are included in the list. An additional function of the application is to track browsing history times. This is done by starting a timer at the beginning of each session, and the application logs the starting and ending time for viewing each page relative to the timer that was started at the beginning of the session. The accuracy the timer is in the range of ±16 milliseconds according to Microsoft MSDN™. The time log is used later to automatically segment the recording. Figure 2.1 shows a

sample of the list given to BrowseHTMList. Figure 2.2 shows the list produced by the program as output that contains time stamps.

BrowseHTMList application is designed and developed to be as generic as possible to benefit the research community running similar data collection sessions. To achieve this goal, it is made open source under a GNU license agreement[4]. Also, it is designed to be easy to customize. The most customizable feature is its ability to host HTML files. HTML provides extensive formatting that is application independent: the font can be changed, pictures can be added, tables can be inserted, and much more. Figure 2.3 shows a snapshot of the application with illustrations about the controls in the application.

Figure 2.1 Part of the list provided as input to the BrowseHTMList application. The first column contains Swadesh list item ids, and the second column contains the HTML page file names

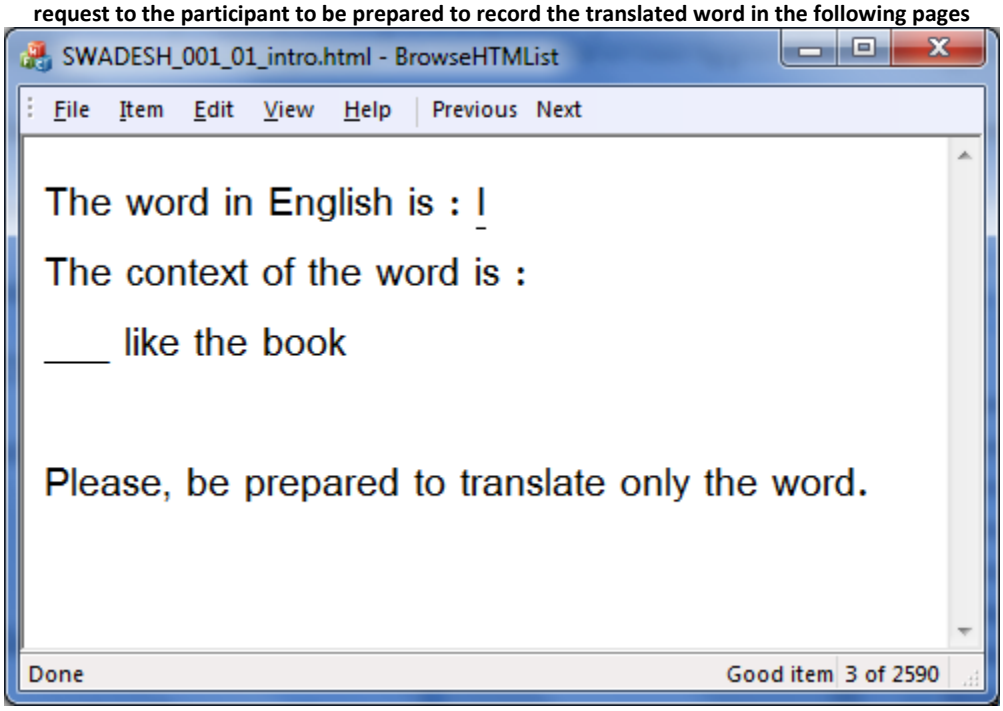
```
no_content_intro_SWADESH_012          SWADESH_012_01_intro.html
SWADESH_012_eng_utter_01              SWADESH_012_02_utter_1.html
SWADESH_012_eng_utter_02              SWADESH_012_03_utter_2.html
SWADESH_012_eng_utter_03              SWADESH_012_04_utter_3.html
no_content_other_var_SWADESH_012      SWADESH_012_05_other_varieties.html
SWADESH_012_var_utter_01              SWADESH_012_02_utter_1.html
SWADESH_012_var_utter_02              SWADESH_012_03_utter_2.html
SWADESH_012_var_utter_03              SWADESH_012_04_utter_3.html
```

Figure 2.2 Part of the list produced as output from the BrowseHTMList application. The first column contains Swadesh list item ids, the second and third columns contain the timestamps in milliseconds of browsing the page relative to the time of loading the list

```
no_content_intro_SWADESH_012          998.406000     1006.986000
SWADESH_012_eng_utter_01              1006.986000    1008.780000
SWADESH_012_eng_utter_02              1008.780000    1011.027000
SWADESH_012_eng_utter_03              1011.027000    1013.289000
no_content_other_var_SWADESH_012      1013.289000    1056.423000
SWADESH_012_var_utter_01              1056.423000    1059.231000
SWADESH_012_var_utter_02              1059.231000    1064.410000
SWADESH_012_var_utter_01              1064.410000    1069.090000
```

[4] BrowseHTMList is downloadable at: https://browsehtmlist.codeplex.com/

Figure 2.3 Snapshot of BrowseHTMList application when the participant is browsing the first item of the Swadesh list. The menu bar shows controls to move to the next page and to move the previous page. The status bar shows the number of items the participant is expected to browse. The content page shows the word in English, the context sentence, and a request to the participant to be prepared to record the translated word in the following pages



A data collection session facilitated by BrowseHTMList would start by starting the sound recording device then loading the list of HTML pages to be browsed during the session. This list should be saved in a text file in the same folder where the HTML pages are located. The "File → Open" menu item loads the list and instantiates a timer that will be used to track the time, loads the first page in the list, and specifies the number of pages that will be browsed in the session in the status bar of the application along with a sequence number of the currently browsed page.

After loading the list, I play a beep by the "Item → SyncBeep" menu item. This beep is used to segment the WAV file as discussed in section 2.4.2. Then the participant or the researcher browse through the list by clicking on the menu items "Next" and "Previous". It was found that the mouse clicks generate undesirable noise in the WAV signal. To avoid this noise, I added the functionality of accessing the menu items using keyboard shortcuts where the keyboard was found to generate less noise. The keyboard shortcuts are "Ctrl+N" for "Next", and "Ctrl+P" for "Previous". To add more flexibility, I added three more menu items in addition the "SyncBeep" menu item under the "Item" menu item to allow the user to skip an item, redo an item, or mark an item as a bad item. These are accessed through "Item → Skip" or "Ctrl+S", "Item → Retry" or "Ctrl+R", and "Item → MarkBad" or "Ctrl+B". Skipping an item is useful in cases where the participant is not recording anything related to the HTML page he/she is viewing. In case of a mistake or mispronunciation, the participant can mark the current item as illformed by clicking on the MarkBad menu item, which shows an indicator on the status bar of the application. Then, redo the recording of that item. There is 0.5 second delay added to the transition between HTML pages when the user moves to the next or previous pages. This delay is to enforce a pause by the participants especially in cases when they are asked to repeat the same word three times because they must wait for the next page to load before each utterance. The data collection session ends by playing another beep then stopping the recording device. The application generates a log file about the browsing session and saves it in the same folder that has HTML pages. See Figure 2.2 for a sample of the log file.

2.4.2 Synchronizing the timestamps and TextGrid boundaries

The lab setting consists of two main systems. The first system records the participant's voice. This system involves an acoustic controlled environment, a microphone and a sound recording device, to be referred to as the recording system. The second system is used to control the flow of the data collection. It presents instructions and stimulus items to the participant, and keeps track of the time it takes in each stimulus item, to be referred to as the flow system. The second system is facilitated by the BrowseHTMList application.

The recording system generates audio recordings in a WAV file while the flow system generates Swadesh list item IDs and timestamps. Linking the timestamps provided by BrowseHTMList within the flow system to the WAV signal of the recording system requires synchronizing the two systems. The synchronization is achieved by having one system generating a signal that is detected by the other system at exactly the same moment. The signal is a sound generated by the flow system and recorded by the recording system. The sound signal is designed to be easy to detect in the WAV signal. It is a beep that the flow system plays within the BrowseHTMList application, this beep is referred as SyncBeep.[5] As mentioned earlier, a menu item is added to the BrowseHTMList application to play the SyncBeep at the beginning of the recording session. Then a Praat script is used to synchronize the timestamps provided by the flow system with the WAV signal. The synchronization provided a perfectly aligned TextGrid in cases where the recording session was short. In long recording sessions where the flow system was operated by the personal laptop of the researcher[6], there was a consistent trend where the beginning of the TextGrid is perfectly aligned while later interval boundaries were more shifted. This consistent pattern seems to be due to inaccuracy of the time tracking hardware in that

[5] The SyncBeep is created using http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/create_waveforms.txt [Type – square, duration – 1 sec, sampling rate 16000, F0 450, amplitude 1, triangle skewness 50]
[6] The researcher's laptop is Dell© Vostro 3500.

particular machine. The shift did not appear when the flow system was operated on another machine. To resolve the problem of misaligned TextGrids, I added the capability of the flow system to generate two SyncBeeps. The first SyncBeep was played at the beginning of the recording session and the second SyncBeep was played towards the end of the recording session. The following equations show the method to calculate the timestamps to set the boundaries of the TextGrid segments based on the time the two SyncBeeps generated by the flow system and detected by the recording system.

x: the input timestamp as indicated by the log generated by the BrowseHTMList application.
b: the difference between the time logged for the first SyncBeep and the time the SyncBeep is actually located in the WAV file
a: the time multiplication factor calculated based on the first and second SyncBeeps
y: the output time to be used to mark the TextGrid boundaries.

$$y = ax + b$$

$$a = \frac{(SyncBeep2\ in\ the\ wav\ signal - SyncBeep1\ in\ the\ wav\ signal\ )}{(SyncBeep2\ in\ log - SyncBeep1\ in\ log)}$$

2.5 Data segmentation

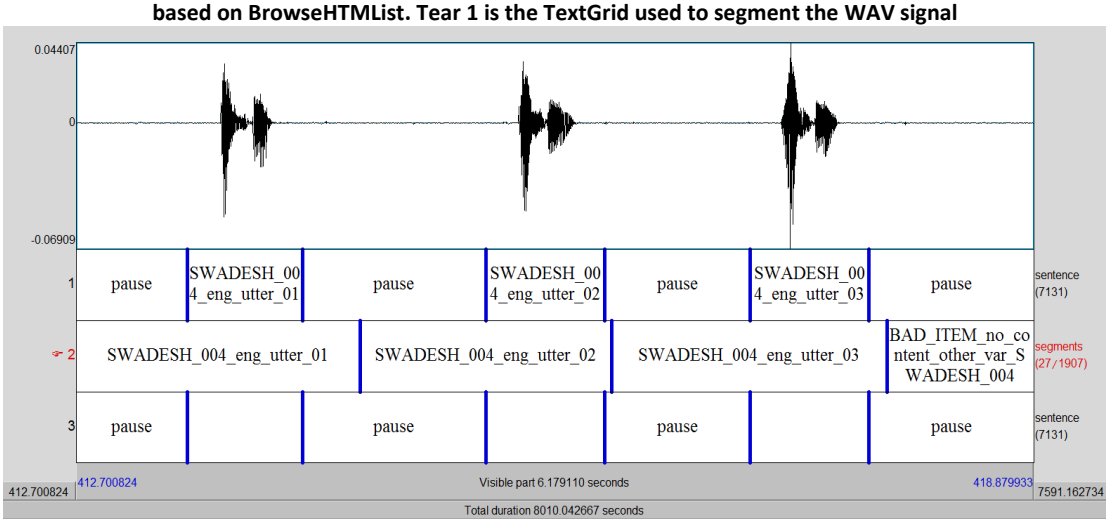
Each data collection session resulted with a sound recording and a log file containing item IDs and timestamps generated by BrowseHTMList. The timestamps were synchronized using the procedure described above. After that, a TextGrid was generated where each line in the log file was associated with its relevant part of the WAV signal. To improve the quality of the alignment, I ran a Praat script that marks the pauses in the WAV signal in a TextGrid.[7] There are

[7] This step is achieved with the help of mark_pauses.praat. A script developed by Mietta Lennes, available at: http://www.helsinki.fi/~lennes/praat-scripts/public/mark_pauses.praat

two TextGrids at this stage, the first marks the IDs of the Swadesh items in the TextGrid intervals. The second marks the intervals of pauses and speech in another TextGrid. A script is developed to get the utterance IDs of each Swadesh item from the first TextGrid, and attach it to the best matching utterance interval in the second grid. This generates a new TextGrid with Swadesh items marked in the intervals of speech. To review the results, all the TextGrids are combined and a script goes though the potential lexical items and plays them for the researcher to review. Mistakes can be easily fixed by having all TextGrids in one Praat window. A manual review was necessary to check for errors and fix cases where the boundaries were not set correctly or when words are given an incorrect ID. Figure 2.4 shows a sample of a TextGrid.

At the end of this stage, I ran a script to resolve any possible redundancy in interval IDs, then another script to segment the WAV signal based on the TextGrid. Each interval ID was given to the extracted file name. The last step is to allow simpler accessibility by creating an HTML page where all items are listed along with their IPA transcription, Arabic script, and links to the sound files of the three utterances.

Figure 2.4 Snapshot from a TextGrid for the production of Swadesh item number 3 as produced by one of the participants. Starting from bottom up: Tier 3 contains the results of the mark pauses step. Tear 2 is the result of the TextGrid generated based on BrowseHTMList. Tear 1 is the TextGrid used to segment the WAV signal



2.6 Transcription

All words were transcribed in both Arabic script and IPA. Arabic script is used to build the list of words that the participants will be given as words in other varieties. All participants were born and raised in a major city in the Arab world so they are expected to be capable of reading words in Arabic script. IPA script is later used as input to the measures of variation (see Chapters 3-5). To minimize the amount of manual work, I developed a script to generate an IPA transcription based on the Arabic script. Avoiding complications in the process of generating IPA script based on Arabic script was achieved by following a set of guidelines to transcribe in Arabic script.

The primary guideline of the transcription based on Arabic script is to have one-to-many mapping between the Arabic symbols and the IPA symbols. The first guideline causes alterations to way the script is normally used. So, the secondary guideline to have the resulting Arabic transcription as comprehendible as possible to educated native speakers of Arabic. I used only one symbol to represent a glottal stop in Arabic script, the symbol is (ء) which is one of the symbols representing the glottal stop. I also borrowed letters from Persian script to represent some sounds that are not represented in Arabic script, the introduced symbols are گ and چ to represent /g/ and/ t͡ʃ/ respectively. In addition, I used the diacritic *sukuun* (used to mark the absence of a vowel in standard Arabic script) to the mark short mid vowel /ə/. All long vowels are transcribed as ا, و and ي. Then, each letter in Arabic is mapped to the corresponding IPA symbol according to Table 2.2. و and ي are considered glides if they are preceded or followed by a vowel and transcribed as *w* and *y* respectively. Otherwise, they are transcribed as long vowels U and I respectively. Some sounds vary from one variety to another; these are mainly ج and the word final *taa marbuuta* ة. They are mapped depending on the participant's dialect. Note that the

Egyptian variety does not have the sounds ʒ and ʤ so the standard Arabic ج is used to represent the sound g, this increases the comprehensibility of the resulting script. After the auto conversion from Arabic to IPA, all words were manually reviewed and corrected in cases of mistakes. As mentioned earlier, Appendix A contains all the words of the Swadesh list form all participants.

MSA's vocalic system contains three vowel categories. Each of the three categories consists of short and long vowels. Short vowels are represented by diacritics and long vowels are represented by letters in the orthography of the language. The number of vowel categories in the spoken varieties is larger. Quantifying the variation between varieties of different vocalic systems is problematic because the variation depends on the granularity of defining vowel categories. Providing a fine-grained representation of vowels that captures, for example, the contrasts between vowels due to the presence or absence of emphatic consonants leads to having a large amount of variation that is derived from having different vowels; however the vowels might be close phonetically. To resolve this problem, I provide two measures of pronunciation variation. One based on a small number of vocalic categories and the other based on the formant frequencies of the vowels in each utterance. MSA is excluded from the measure of pronunciation variation based on the formant frequencies because of the lack of the acoustic data. At the other level where the variation is based on a small number of vowels, the representation of vowels in the spoken varieties should be made comparable to what we know about MSA's vocalic system. One might think that it is better to limit the number of vowel categories to three in an effort to have the same number of vowels in the spoken varieties and in MSA. However, the existence of mid vowels, including schwa, in the dialects makes the problem more complicated. One approach to solve this problem is to set the number of vowels to four in the dialects. So, we have three vowels that are considered similar to the vowels in MSA and a mid-vowel. This is not to

claim that the spoken varieties have only four vowels in their vocalic inventories; most of them have more. The use of four vowels only is justified because we are comparing against MSA and because we have a more fine-grained representation of vowels based on the first and second formant frequencies at a level of comparison where MSA is not included.

The conversion of the consonants in MSA from orthographic letters to IPA is based on the researcher's knowledge of the language and based on the available references of Arabic. The pronunciation of *ji:m* (ج) is considered to be a voiced alveolar affricate *ʤ* (Holes 2004, p. 58).

Table 2.2: Mappings of Arabic letters to IPA letters

Arabic Script	IPA	Arabic Script	IPA	Arabic Script		IPA	
ء	ʔ	ض	dˤ	Variety dependent conversions	ة	ə	**For LA**
ب	b	ط	tˤ			a	**Otherwise**
ت	t	ظ	ðˤ		ج	ʤ	**For MSA and GA**
ث	θ	ع	ʕ			ʒ	**For MA and LA**
ح	ħ	غ	ɣ			g	**For EA**
خ	x	ف	f	Long vowels and glides	ا	A	
د	d	ق	q		و	w	**If preceded or followed by a vowel**
ذ	ð	ك	k			U	**Otherwise (long back high vowel)**
ر	r	ل	l		ي	y	**If preceded or followed by a vowel**
ز	z	م	m			I	**Otherwise (long front high vowel)**
س	s	ن	n	Short vowels	◌َ	a	**low vowel**
ش	ʃ	ه	h		◌ُ	u	**back high vowel**
ص	sˤ	گ	g		◌ِ	i	**front high vowel**
چ	ʧ **Voiceless postalveolar affricate. Also written as / t͡ʃ/**				◌ْ	ə	**mid vowel**
					◌ّ	+	**germination for consonants**

2.7 Predicting vowel landmarks

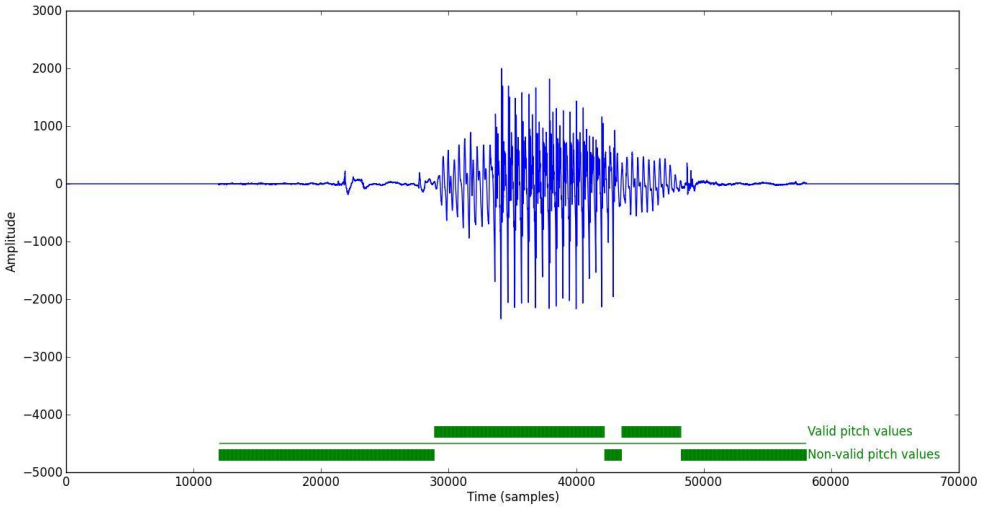
Measuring the formant frequencies for vowels starts by locating a landmark where the formant frequencies are to be sampled. I developed an algorithm to predict a landmark for vowels in the acoustic signal using the IPA transcription as one of the parameters. Vowels are expected to be in the syllable nucleus position and they are expected to be the loudest phonetic segments in the acoustic signal. Mermelstein (1975) developed an algorithm to segment the acoustic signal into syllables based on loudness maxima and minima. The loudness, as he defined it, is a time smoothed and frequency weighted summation of the energy content. De Jong and Wempe (2009) developed an algorithm to detect syllable nuclei in an effort to measure speech rate. Their algorithm is based on locating intensity peaks that are preceded and followed by dips in intensity. Following the same principle, the developed algorithm locates the vowels based on loudness. The inputs to the algorithm are the acoustic signal, the IPA transcription of the word, and the approximated average formant frequencies of the vowels for the speaker. The output is a list of vowels that existed in the input IPA transcription and the predicted landmark for each. A value less than zero is assigned to the landmark when the algorithm fails to locate the vowel. The availability of the IPA transcription is an additional clue that Mermelstein (1975) and De Jong and Wempe (2009) did not have. There are three main benefits of having this extra input: the exact number of vowels, and therefore the number of loudness peaks, is known; the vowels and the approximate values of the formant frequencies are also known – this is given as input to the algorithm; and the number of voiced segments is known, so we can map each vowel to its corresponding voiced segment.

The process of predicting vowel location is divided into four stages using heuristics to automatically locate vowel landmarks with significantly higher than chance accuracy (see Table

2.5). The first stage limits the range of the location prediction by mapping the relevant voiced segment in the IPA transcription to the corresponding voiced segment in the acoustic signal. Identifying voiced segments in the acoustic signal starts by calculating the value of F0 every 1ms, which is a shorter interval than the shortest possible pulse period. Consecutive values of F0 that are within the range of valid values for pitch are considered voiced segments.[8] Supplementing this, consecutive voiced phones in the IPA transcription are also considered voiced segments. If the number of voiced segments in the IPA transcription equals the number of voiced segments in the acoustic signal then the location of voiced segments of the IPA transcription are mapped one-to-one to the corresponding voiced segments of the acoustic signal. If the number of voiced segments in the acoustic signal is more than the number of voiced segments in the IPA transcription then I assume that there are some voiced segments in the acoustic signal that are divided into more than one segment. This issue is resolved by repeatedly merging the smallest voiced segment in the acoustic signal to the closest voiced segment to it until we reach an equal number of voiced segments. Figure 2.5 shows an example of merging two voiced acoustic segments. This problem is apparently due to having some parts of the voiced segment where the calculation of the pitch did not provide a valid value. Which in turn, could be due to creakiness or some distortion in the acoustic signal.

[8] This task is accomplished by running the Praat command: `To Formant (burg)… 0 5 5000 0.025 50.` The threshold for pitch is set to 170 Hz, values more than the threshold are not considered valid pith values. Calculating the pitch using other techniques or having another dataset might result with another threshold.

Figure 2.5 The utterance "baʕdˁ" provided by EA01 in translation for Swadesh item number 20. The utterance is articulated as two voiced segments, the two voiced segments are merged



When there are more IPA voiced segments than acoustic voiced segments, the solution is somewhat more complicated. This could happen if some of the voiced segments were devoiced in some contexts or if some of the unvoiced segments were voiced in a context of voiced segments. The solution is to predict the devoiced segments and ignore them from the IPA transcription. If the mismatch in the number still exists, I force merge the voiced segments in both sides. The force merge accounts for cases of voicing a voiceless phoneme in context of voiced phones. After closely considering the dataset in hand, I composed four phonological rules to account for the cases of devoicing of phonetic segments at word boundary positions. This strategy resolved most of the problems in the dataset. Nevertheless, different languages or datasets might have different rules or different ordering of rules. The rules in the order used for the Arabic dataset in this study are as follows:

1. Ignore a vowel between two voiceless stops in the word initial position. This rule accounts for devoiced vowel between two voiceless stops in word initial position. In such cases, vowels are predicted to be voiceless. Figure 2.6 shows an utterance where a schwa between two voiceless stops was devoiced in word initial position. This caused the acoustic signal to have only one voiced segment while the IPA transcription indicated two. The algorithm ignores the first voiced segment in the IPA transcription and matches the remaining voiced IPA segments to the corresponding segments in the acoustic signal.
2. Ignore a voiced consonant or vowel after a voiceless consonant in the word final position. This rule accounts for devoiced segments of single phoneme after a voiceless consonant in word final position. Figure 2.7 shows an utterance where a word final nasal was devoiced, or barely detectable, after a voiceless stop. The last voiced segment in the IPA transcription was ignored.
3. Ignore a voiced word initial consonant before a voiceless consonant. This rule accounts for a devoiced consonant cluster containing a voiced consonant followed by a voiceless consonant in word initial position. Figure 2.8 shows an utterance where the word initial voiced stop was devoiced before a voiceless fricative. The first voiced segment in the IPA transcription was ignored.
4. Ignore a vowel following a voiceless stop and preceding a voiceless consonant in word initial position. This rule accounts for devoiced vowels after a word initial voiceless stop and before a voiceless consonant. In such case, the vowel is predicted to be voiceless. Figure 2.9 shows an utterance where a vowel after a voiceless stop and before a voiceless fricative was devoiced. Similar to the first three rules, one of the voiced IPA segments is ignored. In this rule, the ignored voiced IPA segment is the first one.

If the number still does not match then merge the voiced IPA segments starting by the first voiced segment and ending by the last voiced segment into one segment. Similarly, merge the corresponding voiced acoustic segments. This ensures that the number of voiced segments in both the acoustic signal and the IPA transcription are equal to one. Therefore, the algorithm never fails to match the number of segments. Figure 2.10 shows an utterance where a voiceless fricative was detected as voiced fricative in context of voiced phones. This is a frequent phonological change that is recovered by the merger rule.

Figure 2.6 The utterance "kətIr" provided by EA01 in translation for Swadesh item number 18. The Schwa is devoiced

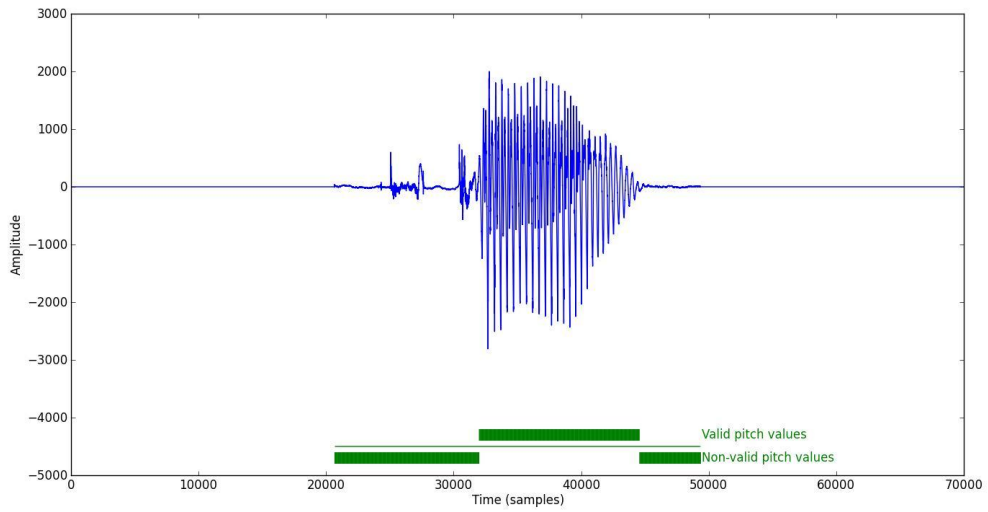


Figure 2.7 The utterance "batʿn" provided by EA01 in translation for Swadesh item number 85. The utterance final nasal is devoiced

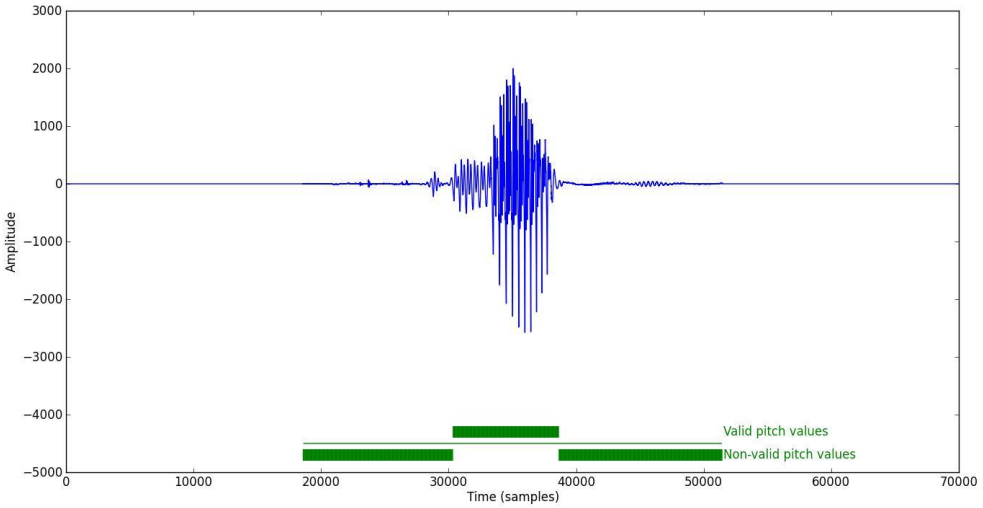


Figure 2.8 The utterance "gsʿIr" provided by GA01 in translation for Swadesh item number 33. The utterance initial voiced stop is devoiced

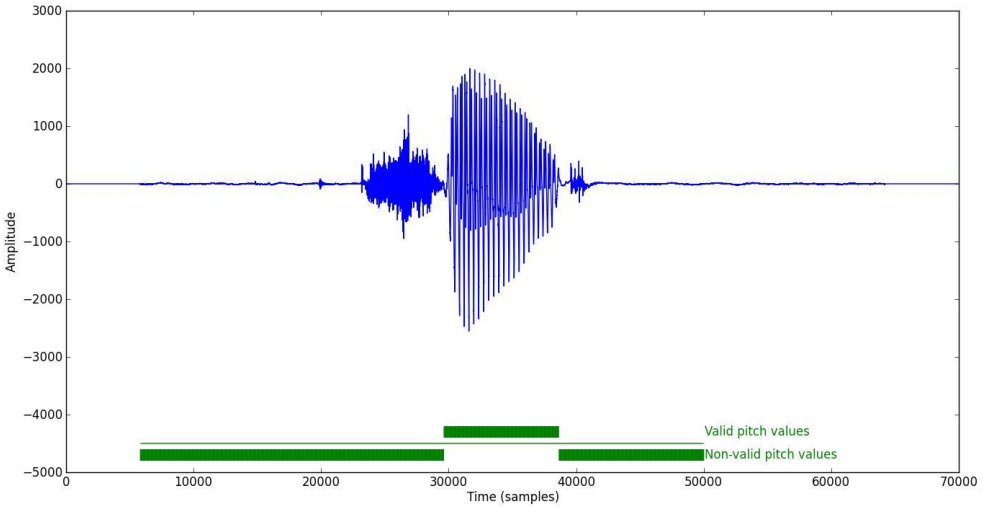


Figure 2.9 The utterance "ʔaħmar" provided by GA02 in translation for Swadesh item number 172. The first vowel is devoiced

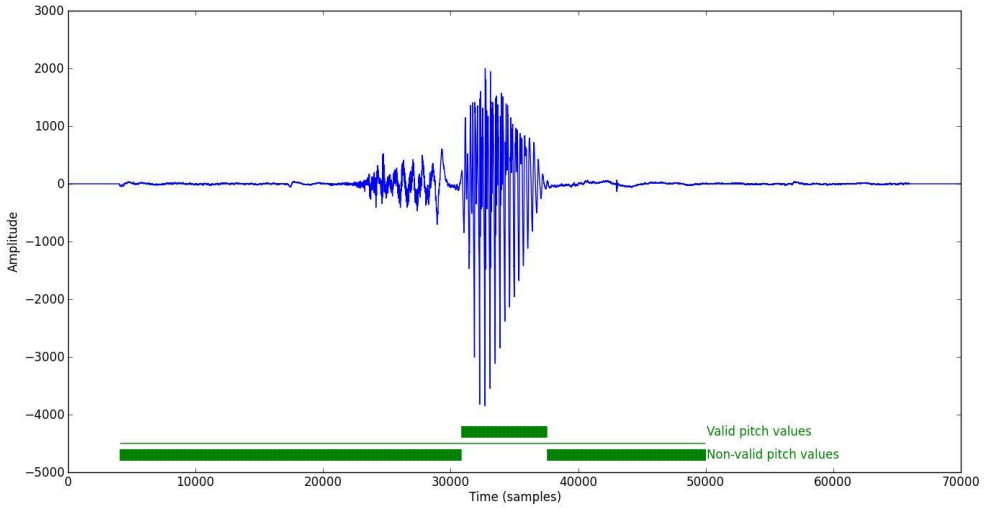
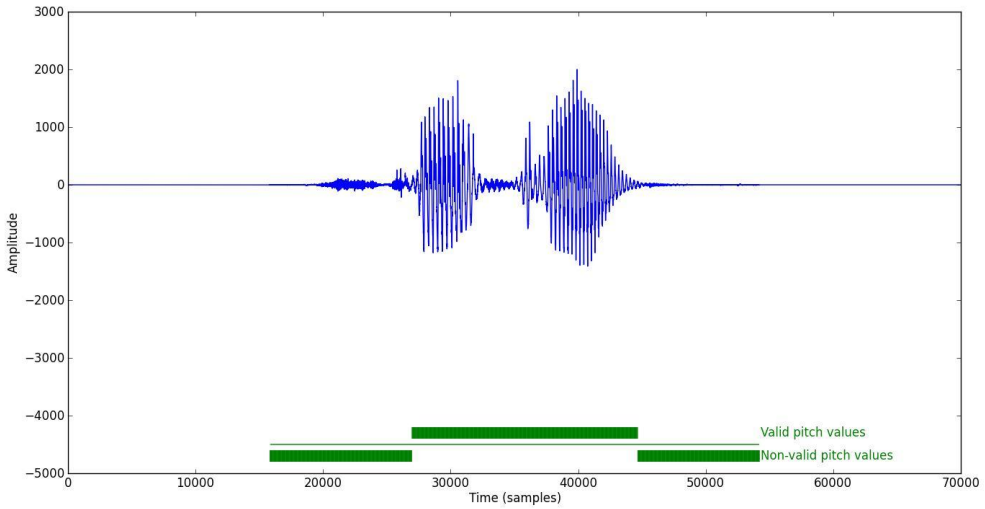


Figure 2.10 The utterance "stafraʁ" provided by EA01 in translation for Swadesh item number 20. The voiceless fricative /f/ is voiced in context of voiced phonemes



The second stage of the algorithm is to evaluate the loudness of the acoustic signal. The first step is to split each voiced segment of the acoustic signal into acoustic units where the loudness is computed and compared, the acoustic units are the pitch periods identified in range of 75 to 500 [9]. For each pitch period in the acoustic signal, the loudness is calculated based on two methods. The first method calculates loudness based on the average amplitude of the absolute values of the sound pressure values, to be referred as the average amplitude method. The second method calculates loudness based on the maximum value of sound pressure minus the minimum value of sound pressure in the pitch period, to be referred as the max-min method. Each method generates a sequence of values representing the loudness of the acoustic signal of the relevant voiced segment. Both of these sequences of values are considered later in the analysis. Figure 2.11 shows the pitch periods and the results of two methods of evaluating the loudness in an utterance containing two voiced segments; the first voiced segment contains two vowels and one vowel in the second voiced segment.

In the third stage, I locate a preliminary landmark for the vowels based on the maxima in the loudness sequences. If the number of maxima in the loudness sequence equals the number of vowels in the corresponding voiced segment, then the location of maxima are set as preliminary predicted vowel landmarks. However, due to natural fluctuations in the acoustic signal, the number of maxima is often more than the number of vowels. In that case, the loudness sequence is smoothed repeatedly with the Simple Moving Average (SMA) algorithm (window size 3) until the number of maxima is equal to or less than the number of vowels. SMA recalculates the value of each point in the sequence as the average of the point itself and points before and after it. So, each value at index *i* is calculated as the average of the values at indices *i-1*, *i*, and *i+1*. In the

[9] This task is accomplished by running the Praat command: `To PointProcess (periodic, cc)... 75, 500`

event that the state of equal number of maxima and vowels was not reached, the prediction based on a loudness sequence fails. For the purposes of this study, SMA provided satisfactory results. However, the vowel prediction algorithm can be potentially improved by experimenting with different smoothing techniques. Figure 2.12 shows the repeated smoothing of the loudness sequences of the utterance plotted in Figure 2.11. For instance, the loudness sequence of the first voiced segment evaluated using the max-min method contained three maxima while the relevant voiced IPA segment contained two vowels. After smoothing the sequence one time, there are still three maxima. Smoothing the sequence again resulted with two maxima that are used as preliminary predicted landmarks for their corresponding vowels.

As mentioned earlier, the preliminary predicted vowel landmarks are set to the maxima of the repeatedly smoothed loudness sequences once a state with an equal number of maxima and vowels is reached. If the preliminary predicted vowel location happened to be in a pulse where the value of the first formant frequency or the value of the second formant frequency is not stable then the algorithm scans for the closest stable pulse located between the minima around the maximum of the preliminary predicted vowel landmark. The stability of a pitch period regarding the formant frequencies is defined by having a standard deviation of the values for both the first and second formant frequencies in the pitch period of less than 50[10]. If a stable pitch period is found then the predicted vowel landmark is set to the center of the closest stable pitch period; otherwise the predicted vowel landmark is set to the preliminary vowel landmark. By the end of this stage, we have predictions from two methods for each vowel with the possibility of failure in one of them or both. Each prediction is evaluated based on the two definitions of loudness in the previous stage and based on the stability of the first and second formant frequencies.

[10] The value of 50 is an ad-hoc number that was set based on trial and error.

Figure 2.11 The utterance "samaka" provided by EA01 in translation for Swadesh item number 45. With illustration of the two methods of evaluating the loudness

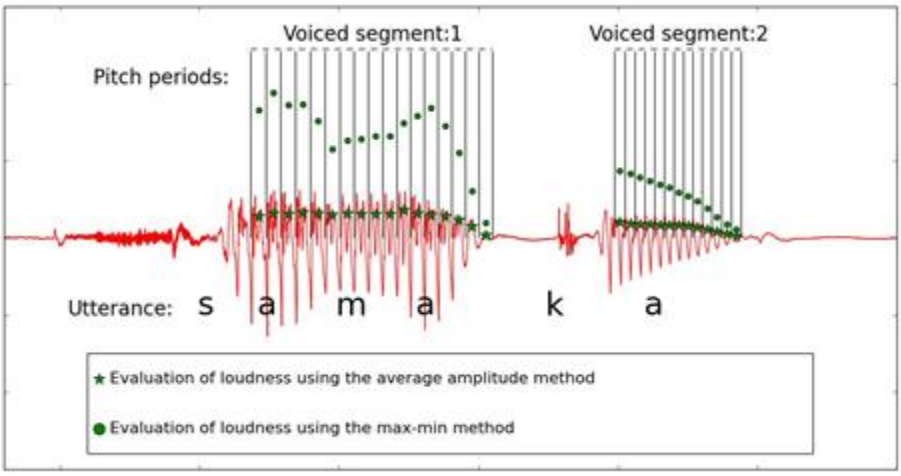
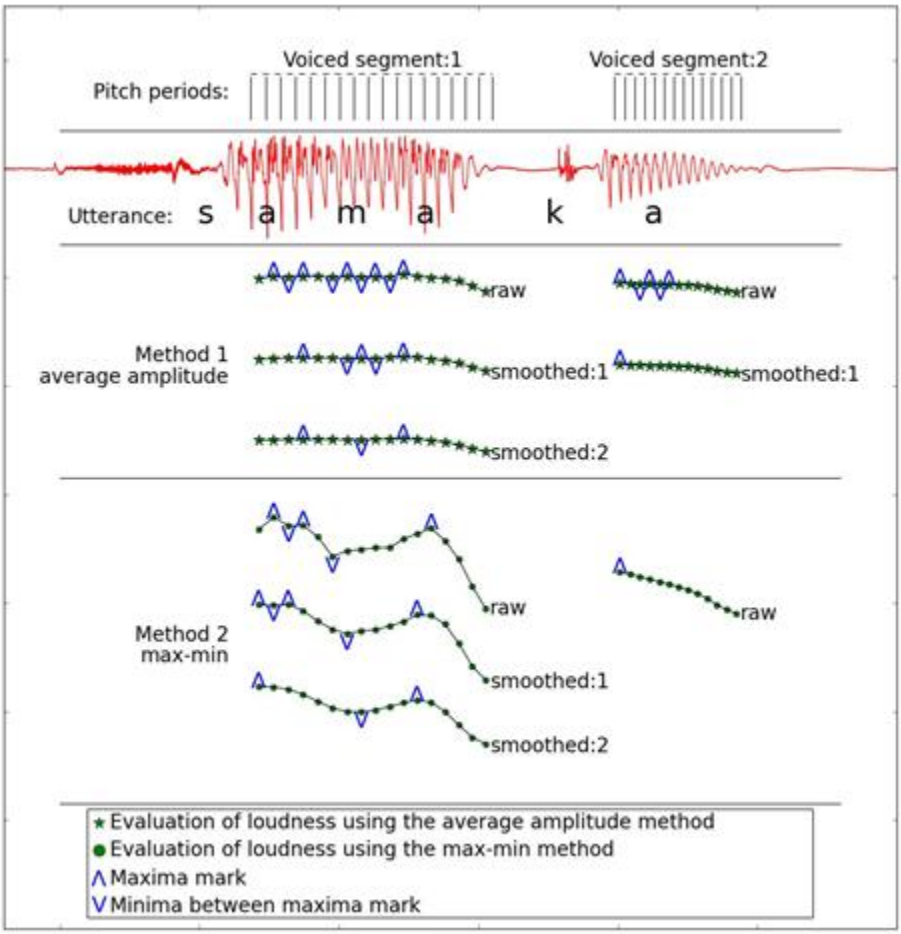


Figure 2.12 The utterance "samaka" provided by EA01 in translation for Swadesh item number 45. With illustration of the prediction of the preliminary vowel landmarks using the two methods



The fourth stage compares the two preliminary prediction values and selects the one that generated values of the first and second formant frequencies closest to the expected values for the first and second formant frequencies of the vowels encoded in the IPA transcription. This stage requires the average values of formant frequencies for each speaker and for each vowel to be estimated. The estimation of the formant frequency values for each speaker and vowel is based on a manually segmented sample of the data. The sample of vowels is a subset of the elicitations of the Swadesh list. From each participant, three distinct words containing productions of each of the four vowels were selected. The total number of the manually segmented vowels is 288 (8 participants × 4 vowels × 3 words × 3 repetitions for each word). For each vowel production, I measured the first and second formant frequencies at the mid-point of the vowel. Then I deleted one of the three repetitions that generated the most distant formant frequencies resulting in 192 vowels. Then I manually deleted 44 vowels because they provided outlying, unreliable values for the formant frequencies. After the last step, the average was calculated based on the remaining 148 utterances. So each vowel's formant frequencies are based on 3 to 6 utterances. Table 2.3 shows the values of the formant frequencies for the vowels for each participant. The estimated values of the formant frequencies were used to select one of the two vowel landmarks predicted by the previous stage of the algorithm. Table 2.4 shows the number of vowels in the data set, the number of predictions produced by each method and the number of predictions selected from each method, as well as the number of vowels for which both methods failed.

To test the accuracy of the prediction algorithm, I randomly selected 132 words from the dataset. All word repetitions are manually segmented so the start and end points for each vowel are known. 737 vowels existed in this data set. The algorithm described above correctly detected

650 vowels of the testing data set: 88.2% of the vowels are correctly detected. The 11.8% failed cases are either due to a failure to smooth the loudness sequence so that the number of maxima equals the number of vowels or due to a misprediction where the predicted vowel landmark is outside the vowel. Table 2.5 summarizes the results of the testing data set. Given the large number of vowels in the study where a manual segmentation of the vowels is not feasible, the accomplished accuracy is considered satisfactory.

Table 2.3: The values of the manually calculated vowel formant frequencies for all vowels for each participant

SPEAKER_ID	VOWEL	formant1	formant2
EA01	a	450	1533
	ə	370	1621
	i	242	2299
	u	297	840
EA02	a	489	1481
	ə	333	1622
	i	272	2083
	u	325	843
GA01	a	450	1170
	ə	426	1224
	i	340	1974
	u	340	1006
GA02	a	612	1219
	ə	454	1331
	i	315	2169
	u	412	766
LA01	a	510	1326
	ə	384	1481
	i	293	2270
	u	340	864
LA02	a	370	1313
	ə	297	1184
	i	273	2018
	u	328	739

Table 2.3: (cont.) The values of the manually calculated vowel formant frequencies for all vowels for each participant

SPEAKER_ID	VOWEL	formant1	formant2
MA01	a	603	1347
	ə	473	1413
	i	302	2154
	u	450	988
MA02	a	463	1319
	ə	374	1312
	i	262	2348
	u	383	938

Table 2.4: Results of the prediction algorithm

	Count	Percentage
Vowel count	9534	100
Cases of vowels predicted to be voiceless by phonological rules	27	0.28
Predictions produced by the average amplitude method	9200	96.5
Predictions produced by the max - min method	9224	96.75
Selected predictions produced by the average amplitude method	6664	69.9
Selected predictions produced by the max - min method	2723	28.56
Both methods failed to predict a landmark	120	1.26

Table 2.5: Results of testing the prediction algorithm

	Count	Percentage
Vowel count in the testing data set	737	100
Cases of vowels predicted to be voiceless	1	0.14
Vowel predicted correctly	650	88.2
Vowel predicted incorrectly	75	10.18
Failed to predict a vowel landmark	11	1.49

2.8 The non-categorical representation of vowels

I obtain a non-categorical representation of vowels by representing each vowel by two numbers derived from the values of first and second formant frequency at the predicted landmark. The objectives of this approach are to provide a more fine-grained representation of the vowels than the one provided by the four categories used in the transcription and to rule out the potential subjectivity of the researcher in deciding what the vowel is in each word. The guidelines of the design of the proposed representation of vowels are (1) to have each vowel represented by two numbers that reflect the place of articulation and the degree of constriction at the place of articulation. (2) To have a considerable amount of the values between 0 and 1. This guideline is used to simplify the way the values are used to calculate the pronunciation variation (chapter 5). The last guideline is (3) to rule out physiological differences among the vocal tracts of the speakers.

I follow an eight-step procedure to achieve the proposed representation of the vowels. The first step is to calculate the values of the first and second formant frequencies at the predicted landmarks for the three repetitions of each Swadesh list item. Therefore, each vowel in each word is represented by three estimations of the formant frequencies. The second step is to delete the prediction that generates the most distant formant frequencies from the average formant frequencies; the averages were calculated for each speaker and for each vowel based on a sample as described in section 2.7. The third step is to eliminate outliers. For each vowel category and for each speaker, I calculated the mean and standard deviation of the distance between the formant frequencies of the vowels and the average formant frequencies. Vowels that are distant more than four times the standard deviation of the distances between vowels and the relevant average vowels were set as outliers. The fourth step is to recalculate the average formant

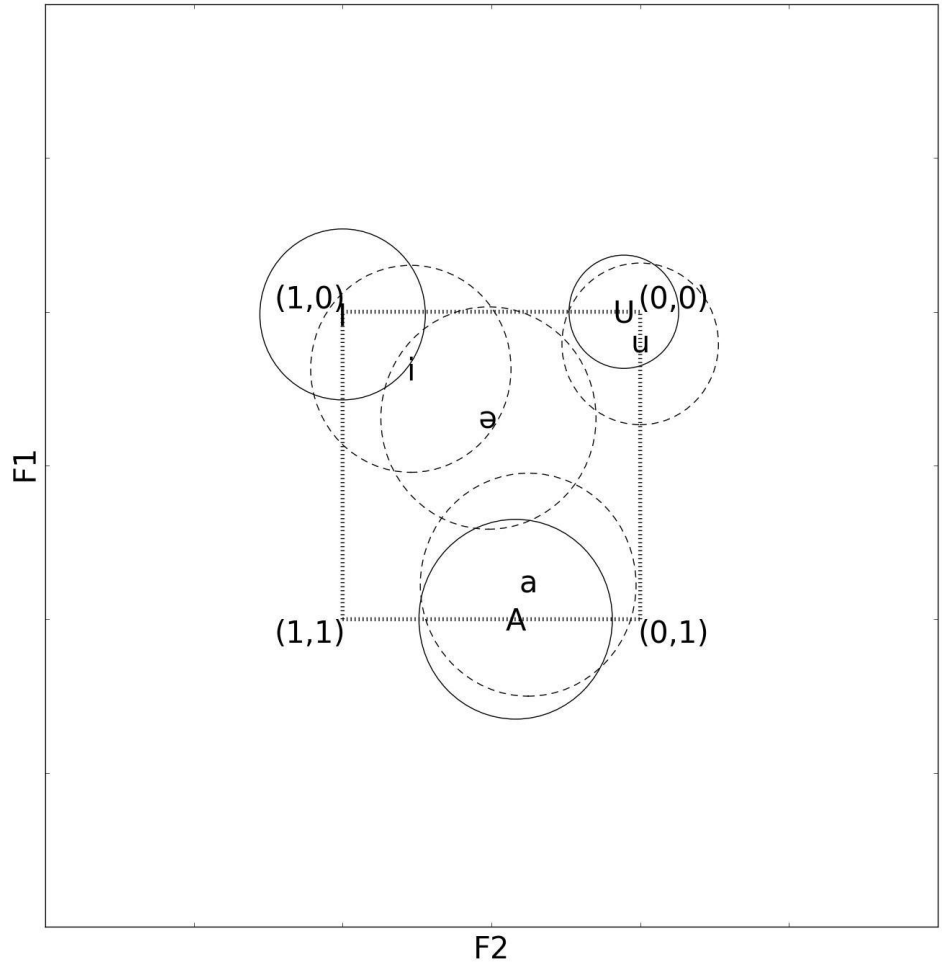
frequencies based on the results of steps one through three and to repeat step two and step three based on the newly calculated averages. This step is motivated because we now have a data set that enables us to achieve a more accurate averages, the averages used in the previous step were based on a relatively small sample as described in Section 2.7. Table 2.6 shows the values of the recalculated averages of the formant frequencies. The fifth step is to assign default values for the formant frequencies of the outliers, the cases of failed predictions and the cases of predicted voiceless vowels as reported in Table 2.4. The default values are the average formant frequencies for each vowel for each speaker. The sixth step is to normalize the values of the formant frequencies to eliminate the physiological differences among speakers. I used Nearey (1978) normalization technique using "Nearey1, formant intrinsic" technique as implemented by Thomas and Kendall (2007)[11]. The seventh step is to scale the normalized values so that a considerable amount of the values is in a range between 0 and 1. This is done by calculating the overall average for the first and second formant frequencies for each vowel across speakers, keep in mind that the previous step normalized the differences between speakers. The smallest and largest averages are scaled to 0 and 1 respectively. The same ratio of scaling applies to all vowels. This method of scaling resulted in 48% of the scaled formant frequencies being in the range of 0 and 1 for both scaled formants. Figure 2.13 shows the positions of the scaled values of the formant frequencies for the vowels. In addition, it shows the dispersion of the values by circles marking one standard deviation around the average values, dashed circles correspond to short vowels and solid circles correspond to long vowels. The corners of the dotted box represent the values ((0,0), (0,1), (1,1), (1,0)). Note that each edge lies on at least one of the averages of either F1 or F2. The eighth step is to encode the vowels in the IPA transcription of the words in the Swadesh list non-categorically based on the values calculated in the seventh step.

[11] The calculation is based on http://ncslaap.lib.ncsu.edu/tools/norm/norm1.php

Table 2.6: The values of the manually calculated vowel formant frequencies for all vowels for each participant

SPEAKER_ID	VOWEL	formant1	formant2
EA01	a	482	1283
	ə	389	1101
	i	317	2037
	u	340	801
EA02	a	451	1391
	ə	373	1638
	i	256	2227
	u	308	923
GA01	a	525	1338
	ə	508	1229
	i	340	1926
	u	370	1047
GA02	a	604	1396
	ə	563	1437
	i	308	2225
	u	382	777
LA01	a	534	1272
	ə	360	1542
	i	290	2270
	u	327	956
LA02	a	478	1383
	ə	279	1554
	i	317	1898
	u	297	909
MA01	a	492	1319
	ə	460	1338
	i	313	1898
	u	380	1010
MA02	a	431	1331
	ə	446	1377
	i	308	2047
	u	349	1041

Figure 2.13 Plot of vowels produced by the speakers. The circles show one standard deviation around the average formant frequencies for each vowel. Dashed circles correspond to short vowels. The corners of the dotted box represent the values ((0,0), (0,1), (1,1), (1,0))



CHAPTER 3

MEASURE OF LEXICAL VARIATION BASED ON THE PERCENTAGE OF NON-COGNATE WORDS

In this chapter, I report on a variation metric based on the percentage of non-cognate words in the Swadesh list. The basic assumption is that the closer the varieties are to each other the more likely they are to have cognate words with the same meaning. A pair of words is identified as cognates if they share the same linguistic origin. Cadora (1979) used a similar method to assess the lexical relationships among major Urban Syro-Lebanese varieties. He used a list of 200 items that consisted of 100 items from the Swadesh list and 100 items from Ferguson-Sa'id's list[12]. Cadora highlighted the possibility of having a pair of cognate words in two varieties with different meanings or with slightly different meanings. He gave the example of the meaning of 'bed' in Damascus and Aleppo as *taxit* and *sariir* respectively. A cognate of *sariir* exists in the variety of Damascus with the meaning of 'crib.' This highlights the importance of specifying the context of the words in the Swadesh list. An example from the Swadesh list used in the current research is the word *fat* that has two senses, as a noun it means the substance *fat* found in human and animal bodies and as an adjective it means *obese*. The first sense can be translated as *simiin* in EA, according to the informant we had. While a cognate of the Egyptian word, *smiin* in LA means the second sense of the word, *obese*. Presenting the participants with only the English words might lead to such confusion, where translation of different senses might be provided. To eliminate confusion, Cadora defined the term of contrastive compatibility as a pair of non-cognate words with the same meaning. In the present

[12] Ferguson-Sa'id's list is not published as far as I know.

study, this problem is resolved by specifying a disambiguating context to each item in the Swadesh list, the context enforces all elicited items to have the same meaning.

Cadora employed the same widely accepted method of using the percentage of non-cognate words to measure lexical distance. He found a correlation between the geographical distance and the lexical distance of some varieties. He divided the Syro-Lebanese varieties into three main groups that reflected their geographical locations. The Lebanese varieties, together with the Syrian variety of Latakia, are categorized as the western group. The dialect of Deir-Ezzor stands alone in the eastern group. The other major Syrian varieties – of Damascus, Homs, Hama, and Aleppo – constitute the central group. He also examined other major varieties of Arabic outside the Syro-Lebanese area and the lexical distance between these varieties and all the Syro-Lebanese varieties in his study (Cadora 1979).

Kessler (1995) used two methods to define cognates. In the first method, called etymon identity, words are defined as cognates if their stem has the same ultimate derivation. In the second method, called word identity, words are defined as cognates only if the words are also cognates at the morphemic level; each morpheme in the word must be cognate in the pair of words. Kessler compared the two methods against previously developed traditional methods to develop dialect maps. The first method seemed to resemble the traditional methods more than the second. In addition to these two basic methods, Kessler developed other metrics that are reviewed with more details in Chapter 4 and Chapter 5.

Gray and Atkinson (2003) used the idea of cognate words to estimate when a set of Indo-European languages diverged from each other. They looked at the shared cognate words between the languages under consideration. It is important to note that the definition of cognate words

they used is different from the one we are using. The main difference is that Gray and Atkinson (2003) exclude cases of borrowing from the list of cognates; as for the present study, words are considered cognates if they have the same linguistic origin, whether by borrowing or genetic inheritance. The difference is justified because the purposes are different. The purpose of the current study is to reflect the degree of mutual intelligibility. On the other hand, Gray and Atkinson's goals were to estimate the divergence time between the languages. The decision of whether a pair of words are cognates or not is a subjective judgment based on the researchers' knowledge of the language. For each entry in the Swadesh list, I assign a unique ID for the set of words that are considered cognates. A table containing decisions of cognates and non-cognates is provided in appendix A.

The design of the lexical variation metric between two varieties in this study is based on the likelihood of a word to be produced as a translation of the Swadesh list item by a speaker to express the meaning implied by the context sentence and the likelihood of a cognate of that word to be also produced by a hearer from the other variety. One of the guidelines for the data collection is to have the participants provide only the words that they would produce when they speak their variety. If both the speaker and hearer would produce a word from the same linguistic origin to express the same meaning then intelligibility is expected to be achieved, which should be implied by a smaller amount of variation. For example, item 39 of the Swadesh list, 'child' in reference to a 5 years old child as the context specifies, is produced as *walad* and *ʕayyil* by EA01 (the first Egyptian participant), and is produced as *tˁifl* and *ʕayyil* by EA02. Since the Egyptian variety is represented by the two participants in this research, the likelihood of the word *ʕayyil* to be produced is 50%, and 25% for each of the other two words. From the perspective of a hearer from the Gulf variety, a cognate of *tˁifl* is available, while no cognate of *walad* or *ʕayyil* was

provided by the Gulf participants. Therefore, 25% of the possible forms are shared in the Gulf variety; the contribution of item 39 in the Swadesh list to the lexical variation metric for an Egyptian speaker and Gulf hearer is an amount of 0.75. Applying the same procedure for all words in the Swadesh list and taking the average of the contribution of each Swadesh list item yield the measure of lexical variation in Arabic varieties.

The current algorithm gives words provided by the speaker(s) equal weights. It might be considered more intuitive to assign bigger weights for more frequent words. For instance, the weights might be derived from a corpus based on the frequency of the words. A corpus to be used in this situation needs to be large enough to contain all the words or at least most of the words of the Swadesh list. While this would be a sound approach, frequency is not considered in the current research because such large corpora unfortunately do not exist for all the Arabic varieties under consideration yet. Therefore, all words are weighted equally for the purposes of the present study.

Table 3.1 summarizes the results of the measure of lexical variation between the varieties of Arabic based on the words elicited by the participants including words they provided based on the English word and English context sentence along with the words they provided based on the data from other participants. Table 3.2 shows the results based on the words that the participants provided before they were shown what other participants provided. This shows the effect of incorporating the extra step of asking the participants about the words provided by other participants. Mainly, this caused the amount of linguistic variation to become smaller for most pairs of varieties to different degrees. I believe that this step is necessary to reliably measure linguistic variation and will be included in the following measure of linguistic variation. Table 3.3 shows the amount of lexical variation between the participants.

Table 3.1 The lexical variation metric between the varieties of Arabic

		Hearer				
		EA	GA	LA	MA	MSA
Speaker	EA		0.17	0.10	0.28	0.14
	GA	0.21		0.14	0.26	0.15
	LA	0.16	0.15		0.27	0.13
	MA	0.30	0.23	0.24		0.22
	MSA	0.19	0.14	0.12	0.25	

Table 3.2 The lexical variation metric between the varieties of Arabic based on words provided only by the English form of the Swadesh list

		Hearer				
		EA	GA	LA	MA	MSA
Speaker	EA		0.19	0.15	0.31	0.14
	GA	0.23		0.15	0.29	0.14
	LA	0.19	0.15		0.29	0.12
	MA	0.32	0.27	0.27		0.23
	MSA	0.22	0.17	0.16	0.29	

Table 3.3 The lexical variation metric between the participants in the experiment

		Hearer								
		EA01	EA02	GA01	GA02	LEV01	LEV02	MOR01	MOR02	MSA
Speaker	EA01		0.04	0.21	0.26	0.17	0.17	0.30	0.32	0.13
	EA02	0.04		0.20	0.25	0.16	0.16	0.29	0.30	0.14
	GA01	0.22	0.22		0.18	0.16	0.20	0.25	0.27	0.13
	GA02	0.23	0.23	0.14		0.20	0.22	0.30	0.30	0.17
	LA01	0.19	0.18	0.15	0.24		0.12	0.27	0.28	0.12
	LA02	0.20	0.19	0.19	0.25	0.12		0.31	0.32	0.13
	MA01	0.31	0.30	0.24	0.32	0.25	0.29		0.05	0.19
	MA02	0.33	0.31	0.26	0.33	0.27	0.30	0.05		0.23
	MSA	0.22	0.22	0.18	0.26	0.17	0.18	0.29	0.29	

EA, LA and GA are closer to each other while MA seems to be more distant from them. As a generalization, we observe from these tables that geographically proximate languages are also lexically more similar based on the lexical variation metric. Considering at the amount of lexical variation between speakers of the same variety, we also observe that the closest pair of participants is the EA speakers since they are from the same city. The MA participants are from different cities in the same country but they are also close to each other lexically. On the other hand, the two GA participants are not as close to each other, compared to the EA and MA participants. The GA participants are from different countries although their cities are close to each other. Similar to the GA participants, the LA participants are also less close to each other. They are from two distant cities located in two countries. As a generalization, based on the limited number of participants in this study, speakers from different countries tend to be linguistically more distant. This generalization should be further investigated by considering more speakers from more diverse geographic distances and from more locations.

Additionally, the asymmetry of the measurement manifests itself when we compare the amounts of lexical variation between some pairs of varieties. For example, when comparing the amount of lexical variation between EA speakers and GA hearers and contrast it with the amount of lexical variation between GA speakers and EA hearers. Such difference in lexical variation could be because the different varieties may have different inventories of lexical items that would facilitate the comprehension of those lexical items in another variety. For example, if a speaker of one variety knows two words for an item on the list, he/she would fully understand a variety that uses only one of those words, but a speaker of that second variety would only understand a speaker from the first variety half of the time.

The amounts of lexical variation between EA speakers and hearers of GA, LA, and MA are less than the amounts of lexical variation between EA hearers and speakers from the corresponding varieties. For example, the amount of lexical variation between EA speakers and GA hearers (0.17) is less than the amount of lexical variation between EA hearers and GA speakers (0.21). This mirrors a pattern of intelligibility we observe regarding communication between Egyptian speakers and members of other varieties; the Egyptian speakers are understood better than they understand members of other varieties. Most of the time, this prompts speakers from other varieties to accommodate for Egyptian speakers. Additionally, we observe from the data that the amounts of lexical variation between LA hearers and speakers of EA, GA, and MA are less than the amounts of lexical variation between LA speakers and hearers from the corresponding varieties. For example, the amount of lexical variation between LA hearers and EA speakers (0.10) is less than the amount of lexical variation between LA speakers and EA hearers (016). This might imply that members of the LA variety are able to understand members of other varieties better than the other varieties understand them.

The results also show that the closest variety to MSA is LA for both the hearers and the speakers, followed by EA and GA: GA is closer to MSA for hearers while EA is closer to MSA for speakers. The farthest from MSA is MA. However, the significance of the differences among some of those measurements is questionable. The next chapter reports on a more fine-grained measure of linguistic variation with more detailed analysis of the reliability of the measure.

CHAPTER 4

MEASURES OF LEXICAL AND PRONUNCIATION VARIATION BASED ON PHONE STRINGS

In this chapter, I consider the phonemic representation to develop a measure of lexical variation and a measure of pronunciation variation by comparing the IPA transcription of the words of the Swadesh list. Comparing all words of the Swadesh list results in a measure of lexical variation that takes into account pronunciation variation (Section 4.1). Comparing only pairs of cognate words results in a measure of pronunciation variation (Section 4.2). Each IPA symbol in the transcription string of a word is considered as an encapsulated unit; the phonetic differences are not taken into consideration. In Chapter 5, I consider the phonetic details at a deeper level of analysis.

The Levenshtein distance algorithm (Levenshtein 1966) provides a measure of sequence similarity. It was invented to measure the similarity between two binary words – a binary word is a sequence of 0s and 1s – for the purposes of detecting distortion of binary data transmitted over a channel. In addition to computer science and engineering, this algorithm has been used in linguistics (Kessler 1995; Heeringa 2004; Serva and Petroni 2008, among others) and biology (Fitch and Margoliash 1967, among others) to measure the similarity between two sequences – a transcription of a word is an instance of a sequence. This algorithm offers a framework for providing a measure of lexical variation that is more fine-grained than the measure of lexical variation discussed in the previous chapter.

Many factors favor the use of the Levenshtein algorithm. First, it is applicable to any sequence, which makes it available to more than one field – linguistics and biology are two

relevant examples. Second, it can solve the problem in a computationally efficient time O(m×n), where m and n stand for the length of the two strings. The major improvements on the efficiency of the Levenshtein algorithm are the ability to approximate the results of the algorithm rather than calculate it precisely.[13] Such improvements of the efficiency of the algorithm are not necessary for linguistic research because the strings under consideration are short, which makes any improvement in the efficiency negligible. Third, it is expandable through the dynamic design of the algorithm. There are two main dynamic aspects of the algorithm: it divides the string into substrings with the substrings being prefixes by default, and it keeps the cost of the basic operations, to be detailed below, independent of the algorithm itself. This specific feature makes the algorithm applicable to linguistic research. I will also propose a new technique utilizing this feature of the algorithm later in this thesis. Fourth, it can be improved to find the best alignment between two strings. This is achieved by keeping track of the places of the insertions, deletions, and substitutions.

The Levenshtein distance algorithm can be defined as a similarity metric that finds the minimum number of insertions, deletions, and/or substitutions needed to transform one string to another. Insertions, deletions and substitutions are called the basic operations of the algorithm. In its most trivial case the cost of each of these operations is set to one. It is also possible to set different costs, and changing the costs might have dramatic effects on the variation metric. In the current chapter, I am setting the cost of basic operations to one, while the next chapter proposes a model of sound representation from which the cost of the basic operations are derived.

[13] For more details see Navarro (2001), Ukkonen (1983), Ukkonen (1985), and Berghel and Roach (1996).

Kessler (1995) was among the first to use the Levenshtein distance algorithm to measure dialect distances. His main objective was to identify the grouping of the Irish Gaelic dialects and to determine the linguistic boundaries between them. Kessler used part of a linguistic atlas developed by Wagner (1958). This part contained a list of 51 concepts represented by 312 different words or phrases. The concepts were presented in narrow transcription based on the IPA standard.

Kessler ran different types of distance metrics that can be divided into two groups. The first group consisted of variation metrics on the lexical level based on the etymon identity and word identity, similar to what was discussed in Chapter 3. The second group of variation metrics considered the IPA transcription and calculated the distance based on the Levenshtein distance algorithm. Within the second group, Kessler introduced a method of phone string comparison. This method was based on the Levenshtein distance with the default cost of the basic operations where all insertions, deletions, and substitutions were set to one. Another technique within the second group was to incorporate the phonetic features in the cost of the basic operations. This technique is called feature string comparison.

Kessler compared the correlation of the variation metrics with the traditional approach of counting the number of isoglosses between dialect sites in a dialect map. The variation metrics based on the Levenshtein distance algorithm outperformed the etymon identity and word identity methods. Within the methods based on the Levenshtein distance algorithm, the phone string comparison method outperformed the methods that considered calculating phonetic differences. Kessler did not conclude that phonetic variation is irrelevant. Rather, he highlighted the importance of further developing methodologies that incorporate phonetic features in the variation metric.

Serva and Petroni (2008) introduced the idea of normalizing the Levenshtein distance between a pair of words over the length of the longer word. This helped ensure that all lexical items are contributing the same weight to the variation metric. The distance between a pair of languages would then be the average of the normalized distances between lexical items. The cost of insertions, deletions, and substitutions were all set to one. The normalization over the length of the longer word generated a distance metric that is less than one for any pair of words which in turn, entailed that the contribution of each lexical item is guaranteed to be less than or equal to one. In other words, all lexical items have the same potential to contribute to the variation metric; assigning weights to lexical items based on frequency was not considered in their study.

Serva and Petroni (2008) used a list of 200 words from 50 languages. Some lists were missing some words, but the maximum number of missing words did not exceed 13. The words were transcribed in English orthography. Based on the known divergence times between two pairs of languages, Serva and Petroni retrieved the divergence times between all other pairs of languages and built a language tree that included the divergence times of all languages in consideration. Previous studies have also built language trees, such as Gray and Atkinson (2003) and Gray and Jordan (2000), but instead of using Levenshtein distance, they focused on the number of non-cognate words as a variation metric.

The use of the Levenshtein distance algorithm has been widely accepted in Linguistics since Kessler (1995). The algorithm was further improved by Serva and Petroni (2008) and Wichmann et al. (2010) by introducing the idea of normalizing the distances. However, their improvements were found to be useful only when comparing distantly related languages. On the other hand, the Levenshtein distance algorithm can be improved by modifying the cost of

insertions, deletions, and substitutions based on sound relatedness or phonetic details. This is one of the main contributions of this thesis, as discussed in detail in chapter 5.

4.1 Measure of lexical variation at the phonemic level

I developed an algorithm to measure the lexical variation based on the IPA transcription of the words of the Swadesh list as transcribed in Appendix A. The algorithm uses the Levenshtein distance algorithm with one as the cost of the basic operations. For each Swadesh list item, the algorithm goes through the words provided by the speaker. For each of those words it finds the closest cognate word provided by the hearer for the same Swadesh list item. The assumption is that the hearer is matching the speaker's word to the closest word in his/her lexicon, and for communication to be successful, both words should have the same meaning i.e. belong to the same Swadesh item. For example, a GA speaker trying to communicate the meaning of the word 'because' (Swadesh item number 206) by using the word *ʕaʃʌn* that exists in his/her lexicon with an EA hearer who has two cognates of this word in his/her lexicon *ʕaʃʌn* and *ʕalaʃʌn*. In such a case, the EA hearer is interpreting the speaker's word to the closest in his/her lexicon which is *ʕaʃʌn*. For this pair of varieties the existence of the word *ʕalaʃʌn* in the hearer's variety does not contribute to the amount of linguistic variation. This component of algorithm also contribute to the asymmetry of the measure because on the other direction of the communication an EA speaker will also be using the word *ʕalaʃʌn* which a GA hearer will match to *ʕaʃʌn* which has a distance of two deletions. Following Serva and Petroni (2008), I normalize the output of the Levenshtein distance algorithm for each pair of words over the length

of the longer word.[14] Then, I normalize over the length of the list. These steps generate a distance that is guaranteed to be less than or equal to one. This ensures that the results of the variation metric are comparable even if some varieties tend to have longer words or if some varieties have longer or shorter lists of pairs words. The algorithm is provided in Figure 4.1.

The results of the measure of lexical variation based on phone strings align with the results of the lexical variation based on non-cognate words provided in the previous chapter. The closest varieties to each other are also the closest geographically: EA, LA and GA. On the other hand, MA seems relatively more distant. The results of this measure also show the two patterns of asymmetry reported by the previous measure. First, the amounts of variation between EA speakers and hearers of GA, LA, and MA are less than the amounts of variation between EA hearers and speakers from the corresponding varieties. Second, the amounts of variation between LA hearers and speakers of EA, GA, and MA are less than the amounts of variation between LA speakers and hearers from the corresponding varieties. It would be interesting to see if those two patterns of asymmetry hold for the pronunciation variation metrics developed in Section 4.2 and chapter 5. The results are given in Table 4.1.

Reliability is an important factor for any measurement procedure. In this study, I provide two tests of the reliability of the algorithm in Figure 4.1. The first test aims to provide a visual realization of the stability of the measure given the size of the Swadesh list. In other words, is the size of the Swadesh list large enough to confidently determine the amount of lexical variation between the varieties under consideration? To answer this question, for each pair of varieties I

[14] Normalizing over the length of the words is an advantage computationally so that each word contributes equally to the computation. However, linguistically there may be reasons to consider developing an algorithm sensitive to word length in future research. One reason is that there does not seem to be an established theoretical definition of *word* Haspelmath (2011), as well as situations where, for example, the average length of words in one variety is shorter than the average length of words in another and this might contribute to the overall difference between these two varieties.

ran a convergence exercise by starting with a list of one randomly selected item. I calculated the amount of variation according to the algorithm described Figure 4.1. I repeated the calculation after growing the list in steps of one randomly selected item. Figure 4.2 shows that the bigger the size of the list is, the more stable the amount of variation between varieties would be. Note that I show a subset of the pairs of varieties in this figure because of the limited space; other pairs of varieties show similar patterns. Based on the patters of convergence, we do not expect that increasing the size of the list would dramatically change the results.

The second test of reliability is based on statistical tools. Assuming that the Swadesh list is a randomly selected sample from the lexicons of the relevant varieties and that the amount of lexical variation between the pairs of words forms a normal distribution, we can use statistical tools to provide a confidence interval for each of the findings reported in Table 4.1. Table 4.2 summarizes the range of 95% confidence intervals for all pairs of varieties. Based on the items of the Swadesh list there is 95% confidence that if we randomly selected similar sized list from the lexicon then the amount of lexical variation between the varieties would fall in the ranges reported in Table 4.2. To assess the closeness of the local varieties to MSA, we focus on the ability of the members of local varieties to comprehend MSA where the speaker belongs to MSA and the hearer belongs to one of the local varieties. Looking at the ranges of the amounts of lexical variation between hearers of the local varieties – EA, GA, LA, and MA – and MSA speakers (highlighted in Table 4.2), we see that MA is more distant relative to MSA than the other local varieties are. Also, GA is closer to MSA than EA is. It could be argued that LA is closer to MSA than EA is and more distant than GA is. However, the ranges overlap and different datasets could provide different results. This variation metric, as the confidence intervals show, did not determine which local variety is closer to MA, as all intervals referring to

MA as either speaker or hearer overlap. On the other hand, it shows that EA speakers are closer to LA hearers than GA hearers are. This might imply that EA is understood by LA better than GA. Moreover, GA speakers are closer to LA hearers than EA hearers are. This also might imply that GA is understood by LA better than EA. On the other hand, the overlapping confidence intervals for the amounts of variation between LA speakers and EA hearers and between LA speakers and GA hearers imply that we are not able to confidently distinguish the closeness of EA and GA hearers to LA speakers based on this measure of lexical variation.

Table 4.1 Results of the lexical variation metric based on the phone string

		Hearer				
		EA	GA	LA	MA	MSA
Speaker	EA		0.32	0.24	0.51	0.36
	GA	0.40		0.27	0.50	0.32
	LA	0.35	0.31		0.51	0.37
	MA	0.52	0.48	0.46		0.51
	MSA	0.38	0.28	0.31	0.52	

Figure 4.1 Algorithm used to measure the lexical variation based on the phone strings of the words of the Swadesh list

```
int Lexical_variation_metric_based_on_phone_string
        (speaker language as LangA, Hearer language as LangB)
{
    Distance_acc = 0
    Word_count = 0
    For each Swadesh_item in the SwadeshList
    {
        Get wordsA_list from LangA that belongs to Swadesh_item
        Get wordsB_list from LangB that belongs to Swadesh_item
        For wordA in wordsA_list
        {
            Get wordB from wordsB_list that is closest to wordA based on Levenshtein dist.
            d = Levenshtein(wordA, wordB)
            d = d / max(length(wordA, wordB))
            Distance_acc = Distance_acc + d
            Word_count = word_count + 1
        }
    }
    Distance = Distance_acc / word_count
    Return Distance
}
```

Figure 4.2 The convergence of the lexical variation metric based on the phone strings, X-axis shows the number of pairs of lexical items in the list. The number of items increases in steps of one. The Y-axis shows the amount of variation based on the algorithm described in Figure 4.1. This figure shows the pattern of convergence for a subset of the pairs of varieties; other pairs show a similar pattern.

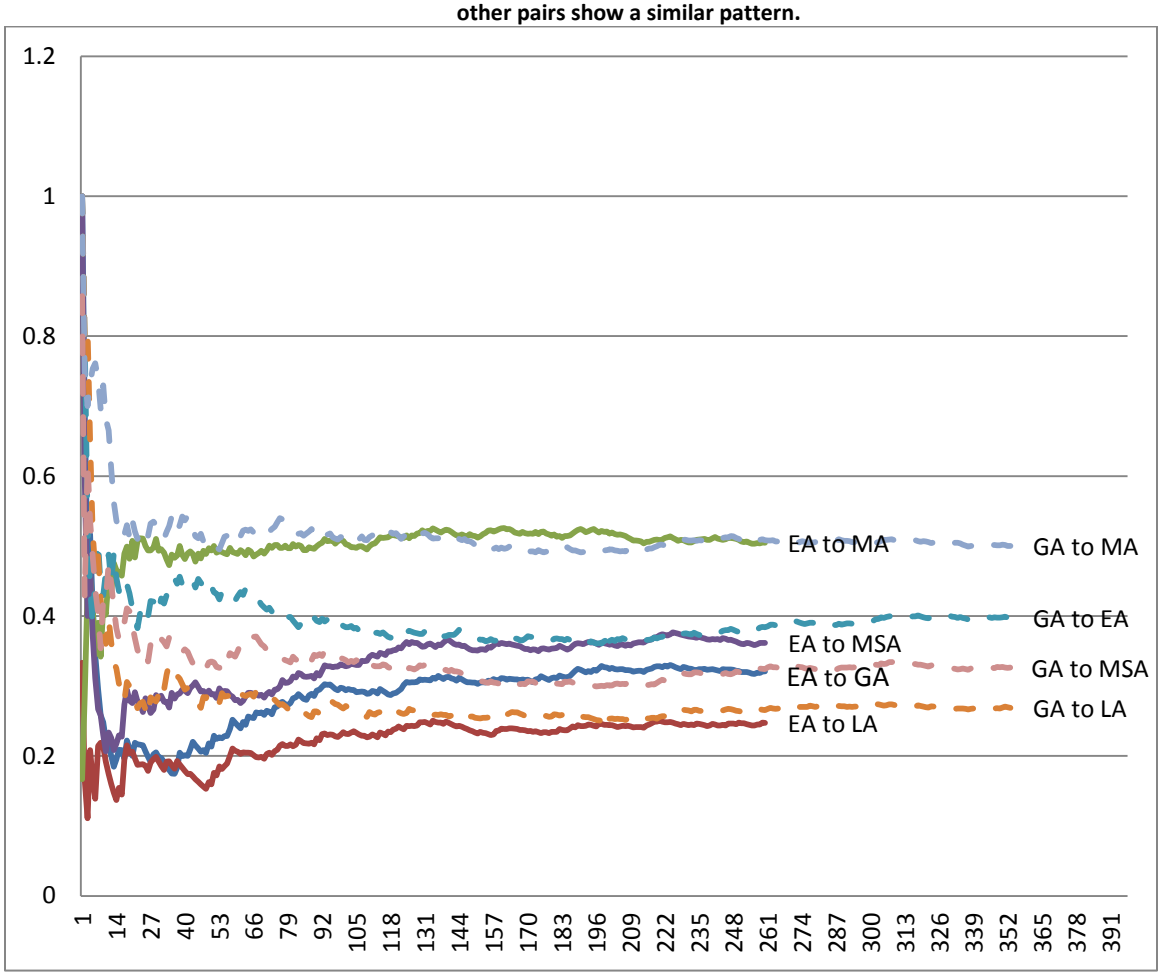


Table 4.2 The range of 95% confidence level of the lexical variation metric between the pairs of varieties

Speaker-Hearer	Degrees of freedom	Mean of normalized distance	Range of 95% confidence interval
EA-GA	257	0.32	0.28 - 0.36
EA-LA	257	0.24	0.21 - 0.28
EA-MA	257	0.51	0.47 - 0.54
EA-MSA	257	0.36	0.32 - 0.4
GA-EA	351	0.4	0.36 - 0.43
GA-LA	351	0.27	0.24 - 0.3
GA-MA	351	0.5	0.47 - 0.53
GA-MSA	351	0.32	0.29 - 0.36
LA-EA	394	0.35	0.32 - 0.39
LA-GA	394	0.31	0.28 - 0.34
LA-MA	394	0.51	0.48 - 0.55
LA-MSA	394	0.37	0.35 - 0.4
MA-EA	272	0.52	0.48 - 0.56
MA-GA	272	0.48	0.44 - 0.52
MA-LA	272	0.46	0.42 - 0.5
MA-MSA	272	0.51	0.47 - 0.55
MSA-EA	269	0.38	0.34 - 0.42
MSA-GA	269	0.28	0.25 - 0.32
MSA-LA	269	0.31	0.28 - 0.35
MSA-MA	269	0.52	0.48 - 0.55

4.2 Measure of Pronunciation variation at the phonemic level

The lexical variation metric reported in the previous section was based on comparing the IPA transcription of pairs of both cognate and non-cognate words. It might be considered problematic to compare the pronunciation of unrelated non-cognate words. But in this case, it is legitimate to do so because the resulting measure – discussed in the previous section – estimates lexical variation based on the phone string across all words, cognate and non-cognates, rather than purely pronunciation variation within cognates as calculated in this section.

In an effort to measure the amount of pronunciation variation, I developed an algorithm similar to the algorithm developed in the previous section except that the comparison of phone strings is limited to pairs of cognate words. This is achieved by incorporating the manually identified cognate words that were developed for the lexical variation metric discussed in Chapter 3. The algorithm is shown in Figure 4.3. The algorithm takes into consideration only pairs of words that are identified as cognates and keeps track of the number of considered pairs of words to normalize over the length of the list.

The results of the measure of pronunciation variation based on the phone strings are given in Table 4.3. Similar to the measures of lexical variation, we still see that EA, LA and GA are closer to each other while MA seems to be more distant from them. Moreover, we still see the pattern of asymmetry for EA speakers: the amounts of variation between EA speakers and hearers of GA, LA, and MA are less than the amounts of variation between EA hearers and speakers from the corresponding varieties. This could imply that members of the EA variety are understood by other speakers better than they understand them. The other pattern of asymmetry for LA speakers is also valid – similar to the lexical variation metric reported in the previous section. The amounts of variation between LA hearers and speakers of EA, GA, and MA are less than the amounts of variation between LA speakers and hearers from the corresponding varieties. This could imply that members of the LA variety understand members of other varieties better than the other varieties understand them. Table 4.4 shows the degrees of freedom, margins of error, and the 95% confidence intervals for the amounts of variation between the varieties according to the current measure of pronunciation variation. Comparing the ranges for the pronunciation variation between MSA speakers and hearers from the local varieties (highlighted in Table 4.4), we notice that GA is the closest followed by LA and EA. Similar to the lexical

variation metric based on phone strings, there is still an overlap for the 95% confidence intervals for MSA-LA and MSA-EA. Moreover, MA is still the farthest to MSA. As for the local varieties, MA hearers are closer to EA speakers than both GA and LA speakers, with no significant distinction between GA-MA and LA-MA. Moreover, due to the overlap of confidence intervals, there is no distinction with regard to the amount of pronunciation variation based on phone strings between MA speakers and hearers of EA, GA, and LA. Similar to the lexical variation metric reported in the previous section, EA speakers are closer to LA hearers than GA hearers and GA speakers are closer to LA hearers than EA hearers. Also, there is no distinction about the closeness of EA hearers and GA hearers to LA speakers. Figure 4.4 summarizes the results of the lexical and pronunciation variation metrics based on the phone string. This plot is provided to make the comparison between the two variation metrics easier. It shows that the results of the pronunciation variation are parallel the results obtained by the lexical variation metric based on the phone string. One area for improvement is to incorporate phonetic features in the measure of pronunciation variation. This is taken up in the next chapter.

Table 4.3 Results of the pronunciation variation metric based on phone strings

		Hearer				
		EA	GA	LA	MA	MSA
Speaker	EA		0.20	0.16	0.34	0.26
	GA	0.23		0.16	0.35	0.21
	LA	0.22	0.19		0.35	0.28
	MA	0.35	0.33	0.31		0.38
	MSA	0.25	0.17	0.22	0.37	

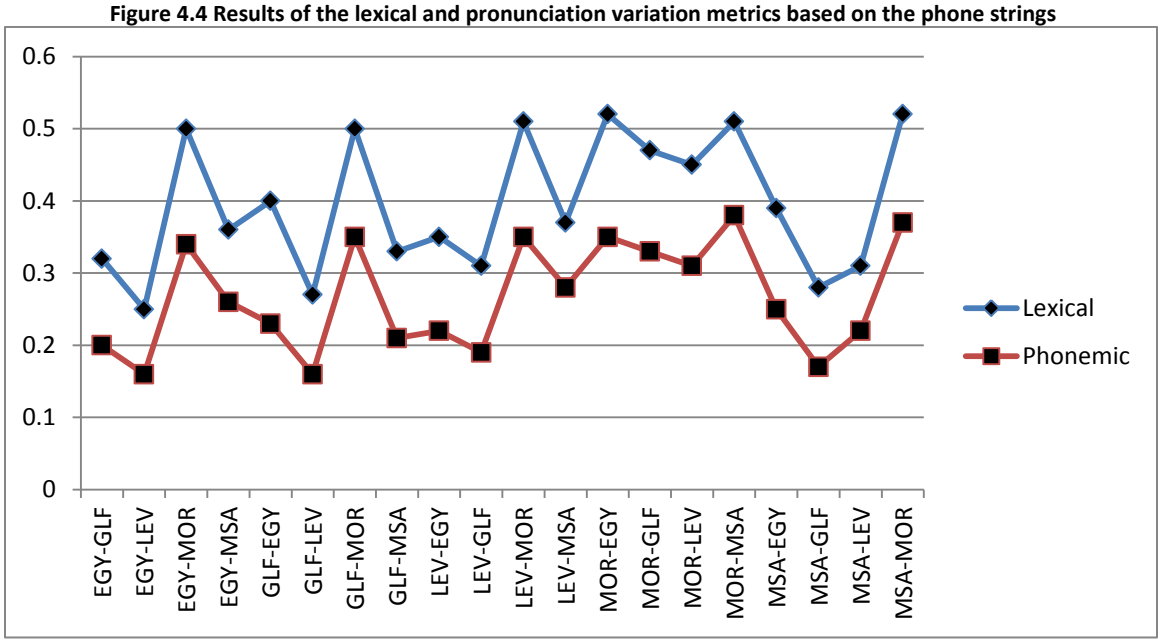
Table 4.4 95% confidence intervals for the measure of pronunciation variation based on phone strings between pairs of varieties

Speaker-Hearer	Degrees of freedom	Mean of normalized distance	Range of 95% confidence interval
EA-GA	206	0.2	0.17 - 0.23
EA-LA	226	0.17	0.14 - 0.19
EA-MA	176	0.35	0.31 - 0.38
EA-MSA	214	0.26	0.23 - 0.3
GA-EA	258	0.24	0.21 - 0.27
GA-LA	288	0.16	0.14 - 0.18
GA-MA	241	0.35	0.32 - 0.38
GA-MSA	282	0.21	0.18 - 0.23
LA-EA	308	0.23	0.2 - 0.25
LA-GA	314	0.19	0.17 - 0.21
LA-MA	266	0.35	0.32 - 0.37
LA-MSA	327	0.29	0.26 - 0.31
MA-EA	183	0.36	0.32 - 0.4
MA-GA	199	0.34	0.3 - 0.37
MA-LA	202	0.31	0.28 - 0.35
MA-MSA	204	0.39	0.35 - 0.42
MSA-EA	205	0.25	0.22 - 0.29
MSA-GA	218	0.17	0.15 - 0.2
MSA-LA	228	0.22	0.2 - 0.25
MSA-MA	188	0.37	0.33 - 0.4

Figure 4.3 Algorithm used to measure the pronunciation variation based on phone strings of cognate words in the Swadesh list

```
int Pronunciation_variation_metric_based_on_phone_string
        (speaker language as LangA, Hearer language as LangB)
{
    Distance_acc = 0
    Word_count = 0
    For each Swadesh_item in the SwadeshList
    {
        Get wordsA_list from LangA that belongs to Swadesh_item
        Get wordsB_list from LangB that belongs to Swadesh_item
        For wordA in wordsA_list
        {
            Get wordB from wordsB_list that is closest to wordA based on Levenshtein dist.
            If wordA and wordB are cognates
            {
                d = Levenshtein(wordA, wordB)
                d = d / max(length(wordA, wordB))
                Distance_acc = Distance_acc + d
                Word_count = word_count + 1
            }
        }
    }
    Distance = Distance_acc / word_count
    Return Distance
}
```

Figure 4.4 Results of the lexical and pronunciation variation metrics based on the phone strings



CHAPTER 5

MEASURES OF PRONUNCIATION VARIATION BASED ON THE MATHEMATICAL REPRESENTATION OF SOUND

This chapter discusses the approach and methodology I follow to develop the measures of pronunciation variation based on phonetic features. As mentioned in chapter 4, one of the favorable features of the Levenshtein distance algorithm is its ability to set variable costs for the basic operations – insertions, deletions, and substitutions. This allows the incorporation of more linguistic details by setting the cost of the basic operations based on phonetic features. For example, the cost of replacing the phoneme /s/ by /z/ should be less than the cost of replacing /s/ by /k/, given that the first pair differs only in voicing while the latter involves more phonetic differences.

As mentioned earlier, Kessler (1995) introduced the use of the Levenshtein distance algorithm to measure linguistic variation. He compared different approaches to compute the distances between Irish Gaelic dialects and compared these with the traditional method: counting isoglosses within a dialect map. Under one of the approaches, he used the Levenshtein distance algorithm with the default cost of one as the cost of the basic operations. Under another approach, he incorporated differences in phonetic features to calculate the cost of the basic operations. He used a set of twelve phonetic features – nasality, stricture, laterality, articulator, glottis, place, palatalization, rounding, length, height, strength, and syllabicity. The values of each feature were set as discrete ordinal values between 0 and 1, with the exact values being arbitrary. Thus, the cost of replacing one phone with another was calculated as the average of the differences between all phonetic features representing those two sounds. Kessler found that the

simpler phoneme-based method with the default cost of basic operations outperformed the multivalued phonetic features method in comparison to the traditional method of counting the number of isoglosses between dialect sites in a dialect map. According to Kessler, the low performance may be due to the arbitrariness of assigning values to the phonetic features.

Heeringa (2004) also used the Levenshtein distance algorithm to calculate the distance between dialects. His study covered a wide variety of ways to calculate the cost of the basic operations. They are divided into two basic categories. The first category is based on phonetic features and the second category is based on the acoustic representation. Within the first category, there are three phonetic feature systems derived from different studies – one based on Vieregge et al. (1984) and Cucchiarini (1983), one based on Almeida and Braun (1986), and one based on Hoppenbrouwers and Hoppenbrouwers (2001). The cost of insertions and deletions is calculated based on the distance between the phoneme and silence while the cost of substitutions is calculated based on the distance between the pair of phonemes being substituted. The distance is derived from segment representation according to the corresponding phonetic representations (Heeringa 2004, p. 124). The methods using acoustic-based representations did not perform as well as the methods using phonetic features.

The phonetic feature systems that Heeringa (2004) used were similar in principle to what Kessler developed in his 1995 study. They both represent phonetic segments by a set of phonetic features and each phonetic feature is associated with a set of ordinal numbers. The differences between them are related to the number of features and the ordinal values assigned to each phonetic feature to distinguish phonetic segments[15].

[15] See Heeringa (2004) section 3.1 for more details.

What is counterintuitive is that both Kessler and Heeringa found that disregarding all phonetic details and using the default cost of one for the Levenshtein algorithm's basic operations produced better results. This finding does not necessarily mean that discarding phonetics details is better but instead may derive from the way costs were assigned, as suggested by Kessler. Thus, such tools have to be designed carefully and should include information about the patterns of sound change that leads to variation.

Gooskens (2007) compared the correlation between the lexical distance and the degree of mutual intelligibility with the correlation between the phonetic distance and the degree of mutual intelligibility. Gooskens found that mutual intelligibility is more correlated with phonetic distance than with lexical distance. He used Heeringa (2004) as a basic phonetic distance metric.

Kondrak (2003) incorporated a new idea in the Levenshtein distance algorithm. In addition to insertions, deletions and substitutions, he introduced the operation of expansion and compression, where a phonetic segment can be expanded or compressed for a specific cost. For each of the 13 phonetic features that he used, he specified a weight, or what he called the salience of the feature, and whether the feature can be applied to vowels and/or consonants. In contrast to the arbitrary nature of the ordinal values that Kessler assigned to his feature set, Kondrak assigned his ordinal values based on physical measurements where applicable. The physical measurements were taken from Ladefoged (1975). The weights assigned to the phonetic features were not based on any physical measurements. Kondrak compared his algorithm with others in terms of its ability to identify cognate words. The comparison included the methods from Kessler (1995), Covington (1996), Somers (1998), Gildea and Jurafsky (1996), Nerbonne and Heeringa (1997), and Oaks (2000). Kondrak's algorithm outperformed them all.

Following this line of research by Kessler (1995) and others, I use the Levenshtein distance algorithm that these researchers have shown to be applicable to measuring pronunciation variation by incorporating phonetic details. The next step is to design a technique to calculate the cost of the basic operations independent of the researcher's intuitions. To achieve this goal, we need to address the following questions:

1. How is the cost of substitutions based on phonetic features derived? How is the cost of insertions and deletions set?
2. What are the sets of phonetic features to be incorporated in representing phones? How are ordinal numbers assigned to values in the phonetic feature sets? How do we assign weights for the different phonetic features?
3. How do we determine if a set of values and weights of phonetic features are better or worse than another set of values and weights? How do we reach the optimal set of values and weights?

5.1 The mathematical representation of sound

This section formalizes a layer of computational representation of sound that is more abstract than the acoustic representation and more detailed than the phonemic representation. The necessity for the new layer of representation of sound comes from the need for an interface that communicates phonetic features that can derive a measure of phonetic similarity. Such an interface is useful in measuring pronunciation variation. At the more abstract level of sound representation, a phonemic based representation, each phoneme is considered as an entity that hides the phonetic features and the fluctuations of the air pressure produced by a speaker uttering the sound. At the proposed layer of representation – the mathematical representation – phonetic

features are encoded. At the more detailed level of representation, the acoustic representation, the fluctuations of the air pressure are recorded over time. Which does not provide a direct interface to communicate the phonetic features.

As stated in the introduction of this thesis, one of the goals for this study is to enhance our understanding of the components of sounds. I am mainly concerned with addressing two questions: (i) What are the key components of sound? and (ii) To what degree is each component playing a role in measuring the similarities and differences of pairs sounds? Answering these two questions is key for developing a measure of pronunciation variation that is more fine-grained than the measure of pronunciation variation based on phone strings (Chapter 4). In addition to the importance of answering these questions to developing a measure of pronunciation variation, their answers might carry potential improvements to some NLP tasks (Chapter 6). Also, they help us provide an empirical framework to answer the theoretical questions about the components of sounds in phonetics and phonology. That said, the specific focus of this study is the development of a measure of pronunciation variation while leaving the additional potential applications for future research.

The quantifiability of pronunciation variation between two sounds is key to the design of the mathematical representation of sound. If each phoneme is represented as a point in a space, then the amount of pronunciation variation between two phonemes is directly derived from the Hamming distance between the points.[16] Within such a design, we need to find the dimensions of the space and the basic principle behind positioning points in the space. For mathematical

[16] Hamming distance is more applicable than Euclidian distance: the former better reflects the changing phonetic features because the latter allows for diagonal shortcuts, increasingly limiting the effect of each individual dimension as the number of dimensions increases. Hamming distance measures the total number of steps on any axis required to reach one point from the other. In other words, the distance is calculated as if a car were to drive around city blocks to reach its destination rather than as the direct path a bird would fly between the points.

simplicity, we assume that each phonetic feature is an independent factor; therefore each phonetic feature corresponds to a dimension in the space.

5.1.1 The phonetic features for encoding in the mathematical representation of sound

The mathematical representation of sound must encode a set of phonetic features that distinguishes all phonemes in the phonemic inventory of the varieties under consideration. But should not include any extraneous features for reasons of computational efficiency. Two frameworks in phonology inspired the set of features used here: the articulatory phonology framework highlights the importance of articulatory gestures, while autosegmental phonology highlights the importance of phonetic features. A purely articulatory model would complicate what could simply be viewed as a phonetic feature. For example, the emphatic feature in Arabic is expressed by set of articulatory gestures including backing the root of the tongue, sagging of the middle of the tongue, and slight rounding of the lips (Abunasser et al. 2011). The complex set of articulatory gestures can be represented as one phonetic feature. On the other hand a purely autosegmental model would fail to capture the relatedness of sound in terms of place of articulation and manner of articulation in a computationally effective way. Drawing from both phonological frameworks and keeping in mind computational simplicity and efficiency, I propose a hybrid model: each phoneme is represented by one main articulatory gesture, while secondary articulatory gestures are considered to produce phonetic features. The resulting model is a representation of sound that can be used computationally in an effective way. The main articulatory gesture is represented by a place of articulation and the degree of constriction at the place of articulation, which allows efficient comparison of different sounds for this core property. The phonetic features are voicing, nasality, laterality, trill/flap, emphasis, lip rounding,

affrication, gemination, and vowel length. A list of the phonemes for the varieties under consideration along with the details about the encoding is provided in Appendix B. The set of phonetic features being used here are derived from the IPA table and are the minimum features required to encode all sounds in the varieties under consideration; other languages might require additional or fewer features. Developing a universal set of features is most likely possible but not necessary at the current stage; furthermore, it increases the computational complexity of other components of the project (see Section 5.1.3).

As mentioned earlier, each phoneme is represented as a point in a multidimensional space where the coordinates of the point specify the main articulatory gesture and the phonetic features. The first dimension specifies the place of articulation of the main articulatory gestures while the second dimension specifies its degree of constriction. The remaining dimensions correspond to the phonetic features, where each feature has its own dimension. The values for each phonetic feature are set to 0 or 1 depending on whether the feature is manifested in the sound or not.

The values in the first dimension (the place of articulation of the main articulatory gesture) correspond to glottal, pharyngeal, uvular, velar, central-vowel, palatal, post-alveolar, alveolar, dental, labiodental, and bilabial, distributed in the range from 0 to 1 in increasing order. Without a phonetically motivated reason for assigning specific values to each intermediate place of articulation, I am proposing a technique that defines the values of the places of articulation as parameters of the representation of sound that will be represented by calculations specific to each pair of varieties (see section 5.1.3).

The second dimension defines the degree of constriction at the place of articulation of the main articulatory gesture. Following the same guidelines of the place of articulation, the smallest value for the degree of constriction corresponds to stops and the largest value corresponds to the degree of constriction of the low vowel; the full range of values is as follows: stop (0), fricative, approximant, high-vowel, mid-vowel, low-vowel (1). The exact values corresponding to the degrees of constriction are parameters that are defined in the following subsections. In previous studies (Kondrak 2003; Heeringa 2004, among others), the consonant and vowel distinction is represented by two separate categories. However, in the current study, the distinction is derived by a gap between the two categories in the second dimension, which is parallel to how the distinction is physically realized (Stevens 2000) and also allows for the computational model to capture the similarity between consonants and vowels, which, for example, can assimilate to one another and otherwise interact.

5.1.2 Parameters and weights of the mathematical representation of sound

The set of phonetic features encoded in the mathematical representation of sound represent each phone as a point in a multidimensional space where the coordinates of the point encode the values of all features. The range of each dimension is 0 to 1. This design results in a computationally effective method to capture sound relatedness. The phonetic distance between a pair of phones can be calculated as the distance between the points representing them. On the other hand, such a design implies that all phonetic features have equal importance because they all have the same range (0 to 1). This problem is resolved by setting weights for all dimensions. The computational component that allows the dimensions to be scaled by the weights needs to be independent of the variation metric and independent of the researcher's intuitions – derived

based on computational calculation. The scaling factors of the dimensions are referred to as weights in the rest of this thesis. Before the Hamming distance between the points is calculated, the axis for each dimension is scaled based on the assigned weight.

The first two dimensions are multivalued where places of articulation and degrees of constriction are expressed as the values of the relevant coordinates between 0 and 1. The exact values of the places of articulation and degrees of constriction are, similarly, to be determined independent from the variation metric and independent from the researcher's intuitions. The coordinates that define the places of articulation and degrees of constriction are to be referred as the parameters of the mathematical representation of sound. However, we need to set default values of the parameters to be used as the starting point in the process of finding the ultimate values of the parameters for each pair of varieties. The default values of the parameters are assigned in a way that they are equally gapped. Table 5.1 reports the default values assigned to the parameters.

Incorporating the mathematical representation of sound in the calculation of pronunciation variation using the Levenshtein distance algorithm entails that the cost of replacements is to be calculated based on the distance between the pair of points corresponding to the pair of phones in question. The question that arises in this context is the following: What is the cost of the other basic operations, insertion and deletion (to be referred as indel[17])? In the new set up that involves the new method to calculate the cost of replacements, keeping the cost of indels to the default cost is not plausible. It might seem plausible to set the cost of indels to the maximum cost of replacement (or a fraction thereof) which is defined as the most distinct pair of phonemes as measured by the distance between the most distant points. However, I do not

[17] The term indel has been used by Kondrak (2003) and in studies in molecular biology.

believe there is a convincing theoretical or logical motivation for such a decision. A computationally feasible and logically plausible solution is to deal with the problem of calculating the cost of indels in a similar method to that of the weights and parameters. The following section discusses the computational component that calculates the weights, the parameters, and the cost of indels.

Table 5.1: Default values of the parameters

Dimension	Parameter name	Default value
place of articulation	Glottal	0
	Pharyngeal	0.1
	Uvular	0.2
	Velar	0.3
	Central vowel	0.4
	Palatal	0.5
	Postalveolar	0.6
	Alveolar	0.7
	Dental	0.8
	Labiodental	0.9
	Bilabial	1
Degree of constriction	Stop	0
	Fricative	0.2
	Approximant	0.4
	High vowel	0.6
	Mid vowel	0.8
	Low vowel	1

5.1.3 Optimizing weights, parameters, and cost of indels based on their ability to identify cognates

This section reports on a computational component that sets values to the Weights, the Parameters, and cost of Indels, to be referred as WPIs. Kondrak (2009) used a phonetic similarity algorithm to identify cognate words. He compared several phonetic similarity metrics based on their ability to identify cognate words. The intuitive assumption is that a better phonetic similarity metric would result with a better cognate word identification algorithm. Following the same intuitive assumption, a better set of WPIs leads to better identification of cognate words for a pair of languages. I find the optimal WPIs based on their ability to identify cognates (as defined in Chapter 3). We need a computational component that given two WPIs can identify which one is better for a pair of varieties. Based on such computational component we optimize for a better set of WPIs.

The best set of WPIs is the set that is able to identify cognates the most; and hence separates cognates and non-cognates the most. In our case, given the Swadesh list for a pair of varieties, the distances between pairs of cognate words form one distribution, and the distances between non-cognate words form another distribution. A good set of WPIs would result in an average distance between non-cognate words to be higher than the average distance between cognate words. Moreover, the more distant the averages are the better the WPIs would be. The distance between the averages of the distributions and the dispersion of each distribution are the key factors that determine the separation of the two distributions. Thus, the more multiples of standard deviations that separate the averages of the two distributions, the better the set of WPIs is. The formula can be derived as:

- Given a list of words, the Swadesh list on our case, for a pair of varieties.

- A: The distribution of the distances between cognate words.
- B: The distribution of the distances between non-cognate words.
- p: a point between A and B that satisfies both 5.1 and 5.2
- x: the multiplication factor of the standard deviation used in equations 5.1, 5.2, and 5.3; x is referred to as the separation factor
- $p = average(A) + x \times std(A)$ (5.1)
- $p = average(B) - x \times std(B)$ (5.2)

- Solving for x yields
  - $x = (average(B) - averae(A)) \div (std(A) + std(b))$ (5.3)

Optimizing for a higher separation factor by setting different weights for each pair of varieties could potentially result with optimal WPIs for each pair of varieties. Such an optimization problem can be solved by implementing a hill climbing algorithm. The hill climbing algorithm consists of repeating two steps. The first step is to start with an arbitrary solution. The second step is to repeatedly improve the solution by finding a better neighboring solution. The process of trying to find a better neighboring solution is repeated until the improvement of the solution fails. Similarly, I start with randomly selected weights, a randomly selected cost of indels, and default values for the parameters. Then the value of each component of the WPIs is increased and decreased by a predefined step size and the WPIs are evaluated each time by calculating the separation factor. Then we select the neighboring WPIs that produced the highest increase in the separation factor. The process of trying to find better neighboring set of WPIs is repeated until the algorithm fails to increase the separation factor. After this point, the last WPIs are considered to produce a local maximum. Following this

algorithm, a local maximum is found for each randomly selected WPIs. After finding a reasonable number of local maxima or repetitions of local maxima the algorithm stops and assumes that the best local maxima is a global maxima and the corresponding set of WPIs are the optimal set of WPIs.

The high computational complexity of the nature of the problem highlights three considerations to keep in mind in order to make it computationally feasible. The first is the step size. A larger step size is computationally less expensive but might lead to a premature local maximum while a smaller step size could be unrealistically computationally expensive. Given the computational resources and after investigating different values and results, the value of the step size is set to 0.1 in an initial stage. Once a local maximum is found, the step size is set to 0.01 and the process repeats one final time. The second consideration is the range from which the random values of the weights and cost of indels are selected. The range is set to the values between 0 and 5 in steps of 0.1. There are 11 weights and 1 value for the cost of indels, thus in total 12 variables to assign random starting values. For each variable, there are 51 possible points to start with, so the total number of possible values is 51^{12}. The third consideration is when to assume that we have found enough number of local maxima and the largest local maxima generates the optimal WPIs? Ideally, we want to be as certain as possible that we have exhausted most of the local maxima and most likely, the global maximum is one of them. The standard procedure for hill climbing algorithms is to start with a pre-specified number of randomly selected starting points, and assume that the best maxima correspond to the optimal result. However, I could not find such a number that is computationally feasible and effective for all pairs of varieties. Instead, the algorithm is designed to repeat the process of starting with randomly selected starting points and find their relevant local maxima until it has five repeated

local maxima. Then the algorithm stops and assumes that the global maximum is the best maximum found and the corresponding set of WPIs is the optimal solution.

One of the challenges to the algorithm is having the starting, randomly selected values begin in a plateau – increasing/decreasing at least one of the values of the starting WPIs will not have any effect on the separation factor. This problem is solved by decreasing the value of each weight, one at a time, as long as the separation factor is not decreasing. This brings the function to an edge of an incline where it may begin climbing and increase the separation factor. The algorithm used to calculate the WPIs is shown in Figure 5.1.

Figure 5.2 illustrates the calculation of the separation factor for a pair of varieties (EA and GA). The x-axis marks the word number and the y-axis marks the distance between pairs of words. The distances between cognate words are marked by pluses and the distances between non-cognate words are marked by circles. The two dotted horizontal lines mark the averages of the two distributions. The two vertical lines mark one standard deviation below and one standard deviation above the average for each distribution. The point *p* is marked by the solid horizontal line. Figure 5.2 (A) shows the separation of the two distributions given the starting randomly selected values. Figure 5.2 (B) shows the separation after the step that avoids having the WPIs in a plateau. Figure 5.2 (C) shows the first change in the weights in an effort to increase the separation factor, there is a step up for the nasal and emphatic features and a stem down for affricate feature. Next, the algorithm adjusts the places of articulation and degrees of constriction. For this pair of varieties and the initial set of WPIs, the algorithm need 53 steps to find a local maxima, the WPIs of the local maxima is given in Figure 5.2 (D).

Figure 5.1 Algorithm used to optimize the WPIs

```
Pseudocode Levenshtein(wpi_set, wordA, wordB)
{
    returns the distance between wordA and wordB following the Levenshtein distance algorithm
by setting the cost of the basic operations using the given wpi_set
}

Pseudocode separation_factor(wpi_set, LangA, LangB)
{
    cognate_list = []
    non_cognate_list = []
    For each Swadesh_item in the SwadeshList
    {
        Get wordsA_list from LangA that belongs to Swadesh_item
        Get wordsB_list from LangB that belongs to Swadesh_item
        For wordA in wordsA_list
        {
            For wordB in wordsB_list
            {
                d = Levenshtein(wpi_set, wordA, wordB)
                d = d / max(length(wordA, wordB))
                If wordA and wordB are cognates
                    cognate_list.add(d)
                else
                    non_cognate_list.add(d)
            }
        }
    }
    return (average(non_cognate_list)-average(cognate_list))
                /(std(cognate_list)+std(non_cognate_list))
}

Pseudocode optimize_parameters(wpi_set, LangA, LangB)
{
    [variables,parameters] = wpi_set
    do
    {
        step_base_sep_factor = separation_factor(wpi_set, LangA, LangB)
        best_wpi_set = wpi_set
        best_step_sep_factor = 0
        For parameter_dim in
            [parameters.places_of_articulation, parameters.degrees_of_constriction]
        {
            for parameter in parameter_dim
            {
                For all possible values of parameter preserving ordinality
                {
                    wpi_set = [x,parameters]
                    sep_factor = separation_factor(wpi_set, LangA, LangB)
                    if sep_factor > best_step_sep_factor
                    {
                        best_step_sep_factor = sep_factor
                        best_wpi_set = wpi_set
                    }
                }
            }
        }
        step_end_sep_factor = separation_factor(best_wpi_set, LangA, LangB)
    }while(step_end_sep_factor > step_base_sep_factor)
}

Pseudocode get_initial_WPIs_set()
{
    variables = get 12 random values in the range 0, 0.1, 0.2 … 5
    parameters.places_of_articulation = default places of articulation
    parameters.degrees_of_constriction = default degrees of constriction
    wpi_set = [variables,parameters]
    return wpi_set
}
```

Figure 5.1 (cont.) Algorithm used to optimize the WPIs

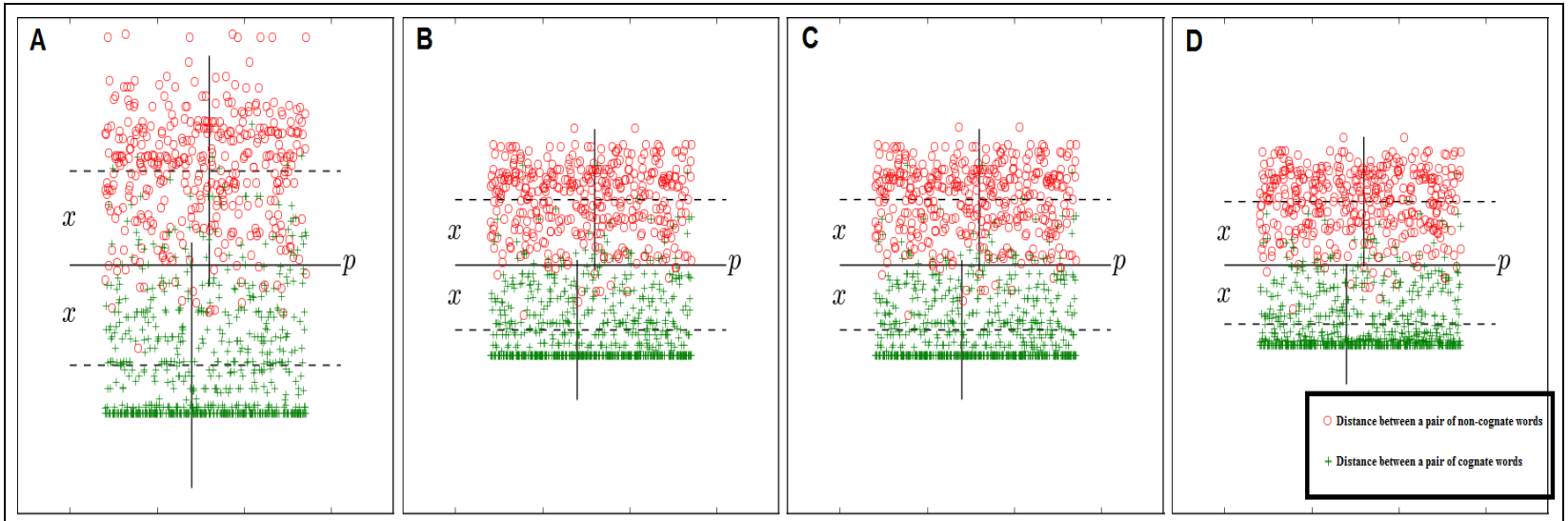
```
Pseudocode climb_the_hill(wpi_set, step_size, LangA, LangB)
{
    [variables,parameters] = wpi_set
    do
    {
        step_base_sep_factor = separation_factor(wpi_set, LangA, LangB)
        best_wpi_set = wpi_set
        best_step_sep_factor = 0
        L = all neighboring variables
        For x in L
        {
            wpi_set = [x,parameters]
            sep_factor = separation_factor(wpi_set, LangA, LangB)
            if sep_factor > best_step_sep_factor
            {
                best_step_sep_factor = sep_factor
                best_wpi_set = wpi_set
            }
        }

        optimize_parameters(best_wpi_set, LangA, LangB)
        step_end_sep_factor = separation_factor(best_wpi_set, LangA, LangB)
    }while( step_end_sep_factor > step_base_sep_factor)
}

Pseudocode move_down_weights(wpi_set, step_size, LangA, LangB)
{
    moved_down = true
    while (moved_down)
    {
        moved_down = false
        [variables,parameters] = wpi_set
        For each variable in variables
        {
            while (variable >= step_size)
            {
                variable = variable - step_size
                new_wpi_set = wpi_set with the new value of variable
                If (separation_factor(new_wpi_set, LangA, LangB)
                    >= separation_factor(wpi_set, LangA, LangB))
                {
                    wpi_set = new_wpi_set
                    moved_down = true
                }
            }
        }
    }
}

Pseudocode OptimizeWPIs(LangA, LangB)
{
    while(number of already seen maxima < 5)
    {
       wpi_set = get_initial_WPIs_set()
       for step_size in [0.1, 0.01]
       {
            move_down_weights(wpi_set, step_size, LangA, LangB)
            climb_the_hill(wpi_set, step_size, LangA, LangB)
            #wpi_set generates a local maxima
       }
    }
    return wpi_set
}
```

Figure 5.2 Illustration of the separation factor



Separation factor(x): 1.630 indel:2.1 place:4.7 constriction:1.4 voice:4.2 nasal:2.3 lateral:2.8 trill_flap:3.3 affricate:2.8 rounded:1 long_vowel:1.8 geminated:2.2 emphatic:2.9	Places of articulation	b:1 l:0.9 d:0.8 a:0.7 e:0.6 p:0.5 c:0.4 v:0.3 u:0.2 r:0.1 g:0	Separation factor(x): 1.852 indel:1.3 place:4.1 constriction:0 voice:2.6 nasal:1.3 lateral:1.8 trill_flap:2.5 affricate:0.1 rounded:0 long_vowel:0.2 geminated:0.9 emphatic:1.5	Places of articulation	b:1 l:0.9 d:0.8 a:0.7 e:0.6 p:0.5 c:0.4 v:0.3 u:0.2 r:0.1 g:0	Separation factor(x): 1.857 indel:1.3 place:4.1 constriction:0 voice:2.6 nasal:1.4 lateral:1.8 trill_flap:2.5 affricate:0 rounded:0 long_vowel:0.2 geminated:0.9 emphatic:1.6	Places of articulation	b:1 l:0.9 d:0.8 a:0.7 e:0.6 p:0.5 c:0.4 v:0.3 u:0.2 r:0.1 g:0	Separation factor(x): 1.957 indel:1.18 place:5.11 constriction:0.81 voice:2.36 nasal:2.36 lateral:1.91 trill_flap:2.36 affricate:0 rounded:0.36 long_vowel:0.14 geminated:0.45 emphatic:0.73	Places of articulation	b:1 l:1 d:0.7 a:0.7 e:0.7 p:0.61 c:0.59 v:0.59 u:0.36 r:0.13 g:0
	Degrees of constrictions	s:0 f:0.2 t:0.4 h:0.6 m:0.8 w:1		Degrees of constrictions	s:0 f:0.2 t:0.4 h:0.6 m:0.8 w:1		Degrees of constrictions	s:0 f:0.2 t:0.4 h:0.6 m:0.8 w:1		Degrees of constrictions	s:0 f:0.12 t:0.57 h:0.8 m:0.8 w:1

The optimal WPIs are calculated twice for each pair of varieties. The first time, the vowels are represented categorically based on the phonetic transcription. In this case long vowels are in three categories and short vowels in four; in the case of short vowels, there is schwa (Section 5.2). The second time, the vowels are represented based on the values derived from the formant frequencies reported in Section 2.8 (Section 5.3).

5.2 Measure of Pronunciation variation based on the mathematical representation of sound

Following the algorithm presented in Figure 5.1, I calculate the WPIs for each pair of varieties. I ran the procedure twice for each pair of varieties to show the consistency of the algorithm in finding the optimal WPIs. Most values are very close to each other, if not exactly the same. The reliability of the algorithm could be enhanced by increasing the number of repeated local maxima required to find the optimal WPIs to a value bigger than five or by having a smaller step size. However, the achieved accuracy is considered satisfactory given the computational resources on hand. The optimal WPIs for all pairs of varieties are provided in Table 5.2. Then, the optimal WPIs for each pair of varieties are considered those that generated a bigger separation factor. See Section 6.2 and Section 6.3 for issues related to the values in this table.

The amount of pronunciation variation is calculated based on the algorithm provided in Figure 4.3 with the cost of each basic operation in the Levenshtein distance algorithm calculated based on the optimal WPIs for the relevant pair of varieties. Table 5.3 summarizes the results. The closest varieties to each other are the closest geographically: LA, GA, and EA. MA is relatively more distant both geographically and based on the current measure of pronunciation variation. As with the previous measures, we still see the two patterns of asymmetry. First, the

amounts of variation between EA speakers and hearers of GA, LA, and MA are less than the amounts of variation between EA hearers and corresponding speakers from the those varieties. Second, the amounts of variation between LA hearers and speakers of EA, GA, and MA are less than the amounts of variation between LA speakers and hearers from the corresponding varieties.

Table 5.4 reports the 95% confidence intervals for the amounts of pronunciation variation reported in Table 5.3. The highlighted rows show the intervals for the amounts of variation between MSA speakers and hearers from local varieties. As mentioned earlier it is more important to show the potential of the members of local varieties to comprehend MSA. GA hearers appear to be the closest to MSA speakers, followed by EA then LA. However, the 95% confidence intervals for those pairs of varieties overlap – the first three highlighted rows in Table 5.4. This means that we cannot confidently determine which of the three varieties is the closest to MSA from the measure of pronunciation variation based on the mathematical representation of sound. On the other hand, there is no overlap for the interval corresponding to MSA-MA with the other intervals for the formerly mentioned local varieties. So, the results of the current measure show that GA, EA, and LA are all closer to MSA than MA.

The closest local variety to MA is EA considering both directions of communication – MA speakers to EA hearers and EA speakers to MA hearers. On the other hand, we cannot distinguish between the measures of closeness for LA and GA to MA due to the overlap of the relevant confidence intervals. Similar to the previous measure, EA speakers are closer to LA hearers than GA hearers are. Also, GA speakers are closer to LA hearers than EA hearers are. As for LA speakers, there is no distinction regarding the closeness of EA hearers and GA hearers to them.

Figure 5.3 summarizes the results for the lexical, pronunciation based on phone strings, and pronunciation based on mathematical representation methodologies. This plot is provided to make the comparison of the three variation metrics easier for the reader. It is not valid to compare the values from different variation metrics directly, but comparing the relative values within each metric is informative.

Figure 5.3 Comparison of three variation metrics

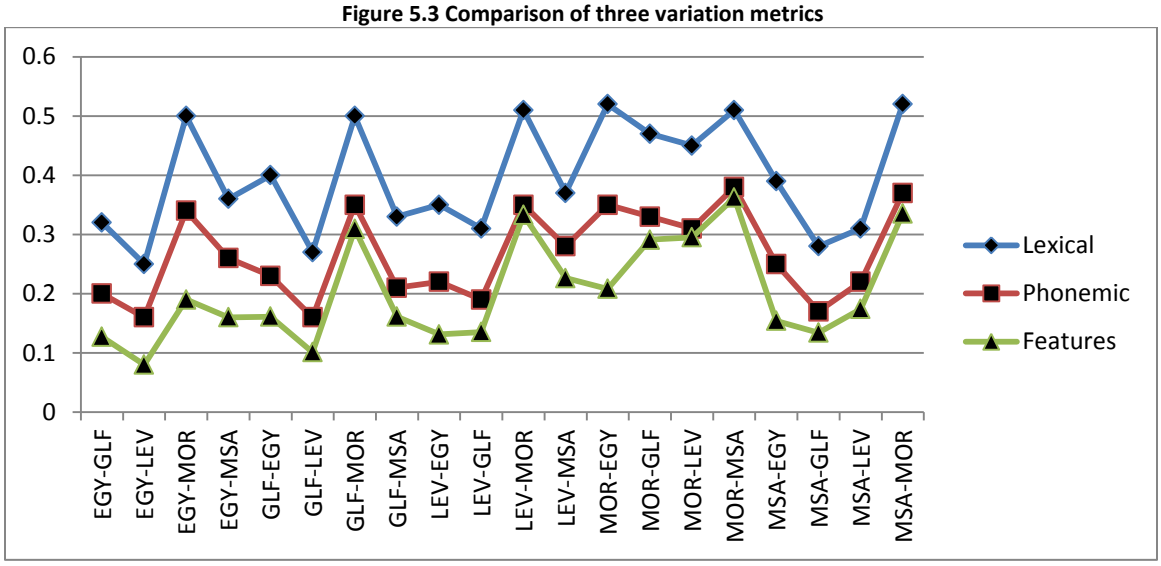


Table 5.2 WPIs calculated based on phonemic representations of vowels

	pair_name	EA-GA	EA-GA	EA-LA	EA-LA	EA-MA	EA-MA	EA-SA	EA-SA	GA-LA	GA-LA	GA-MA	GA-MA	GA-SA	GA-SA	LA-MA	LA-MA	LA-SA	LA-SA	MA-SA	MA-SA
	Trial number	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
	separation_factor	2.077	2.026	2.227	2.212	1.830	1.829	2.298	2.266	1.966	1.960	1.651	1.648	2.324	2.322	1.779	1.775	2.049	2.044	1.772	1.772
	p	0.574	0.559	0.530	0.525	0.600	0.601	0.721	0.667	0.641	0.641	0.817	0.819	0.652	0.645	0.915	1.011	0.715	0.710	0.897	0.888
	cognates_mean	0.162	0.158	0.133	0.136	0.221	0.223	0.178	0.165	0.161	0.168	0.328	0.328	0.169	0.167	0.364	0.410	0.233	0.232	0.384	0.380
	cognates_std	0.199	0.198	0.178	0.176	0.208	0.206	0.237	0.221	0.244	0.242	0.296	0.298	0.208	0.206	0.310	0.339	0.235	0.234	0.290	0.287
	non_cognates_mean	0.972	0.959	0.882	0.875	0.923	0.925	1.275	1.165	1.166	1.154	1.277	1.275	1.258	1.240	1.367	1.491	1.177	1.161	1.400	1.386
	non_cognates_std	0.192	0.197	0.158	0.158	0.176	0.177	0.241	0.220	0.267	0.262	0.279	0.277	0.261	0.256	0.254	0.270	0.225	0.221	0.284	0.281
	numer of local maxima	70	50	21	41	28	47	50	39	70	93	45	48	88	61	51	89	38	39	23	26
Weights	place	5.03	4.99	2.43	2.53	2.03	1.73	5.29	4.56	5.71	5.11	6.67	12.68	7.65	7.56	4.51	4.17	5.73	4.86	10.96	10.8
	constriction	1.99	2	2.03	0.93	0.69	1.33	3.82	2.42	1.54	0.81	1.19	0.25	2.44	1.99	4.31	2.37	1.92	1.77	0.88	0.88
	voice	1.55	1.6	1.58	1.6	1.31	1.27	2.71	2.44	2.4	2.36	1.37	1.36	1.24	1.25	1.24	1.35	1.5	1.46	1.94	1.92
	nasal	1.9	1.91	1.8	1.8	1.58	1.48	2.7	2.45	2.4	2.36	2.31	2.31	2.6	2.6	2.52	2.67	2.38	2.34	2.76	2.7
	lateral	0	0.29	1.6	1.36	0.8	0.96	2.12	0.2	1.96	1.91	1.1	1.98	0.26	1.26	0.74	1.9	0	0	1.68	1.7
	trill_flap	1.8	1.8	1.6	1.18	1.17	1.14	2.7	1.65	1.65	2.36	2.31	2.31	1.79	1.23	2.64	2.89	2.38	2.35	1.98	1.94
	africated	0	0	0	0	0.01	0.01	0	0	0	0	0	0	0	0.45	0	0	0	0	0	0
	rounded	0.26	0.29	0	0	0.17	0.16	0	0	0.27	0.36	1	0.09	0.17	0.15	0	0.06	0	0	0	0
	long_vowel	0.15	0.11	0.02	0.01	0	0	0	0	0.23	0.14	0.09	0.13	0.04	0.04	0.16	0.31	0.02	0.01	0.13	0.13
	geminated	0.81	0.38	0.41	0.41	1.52	1.49	0.56	0.54	0.03	0.45	0	0	0	0	0	0	0	0	0	0
	emphatic	0.59	0.58	0.33	0.12	1.58	1.58	0	0	0.83	0.73	1.25	1.26	2.6	1.87	1.18	2.39	0.69	0.71	1.06	1.05
	indel	0.9	0.9	0.79	0.8	0.79	0.79	1.35	1.22	1.2	1.18	1.31	1.31	1.3	1.28	1.32	1.44	1.19	1.17	1.37	1.35
Places of articulation	glottal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	pharyngeal	0.2	0.15	0.17	0.17	0	0	0.2	0.14	0.2	0.13	0.18	0.08	0.16	0.16	0.03	0	0.08	0.09	0.06	0.06
	uvular	0.28	0.25	0.5	0.41	0.07	0.16	0.2	0.15	0.32	0.36	0.4	0.2	0.5	0.5	0.2	0.2	0.2	0.25	0.31	0.31
	velar	0.69	0.7	0.5	0.5	0.4	0.5	0.5	0.4	0.5	0.59	0.4	0.2	0.5	0.5	0.2	0.2	0.2	0.25	0.31	0.31
	central_vowel	0.7	0.7	0.5	0.5	0.4	0.5	0.5	0.4	0.5	0.59	0.41	0.28	0.5	0.5	0.41	0.56	0.2	0.26	0.38	0.38
	palatal	0.7	0.7	0.5	0.5	0.4	0.5	0.5	0.4	0.51	0.61	0.47	0.31	0.52	0.53	0.47	0.64	0.2	0.27	0.43	0.43
	postalveolar	0.7	0.7	0.5	0.5	0.4	0.5	0.5	0.5	0.7	0.7	0.7	0.7	0.7	0.56	0.6	0.7	0.7	0.7	0.7	0.7
	alveolar	0.7	0.7	0.5	0.6	0.4	0.5	0.61	0.59	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.7	0.7	0.7	0.7	0.7
	dental	0.7	0.7	0.5	0.6	0.8	0.8	0.61	0.59	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.7	0.7	0.7	0.7	0.7
	labiodental	1	1	1	1	1	1	1	1	1	1	1	1	0.79	0.67	0.94	0.98	0.95	0.95	0.94	0.94
	bilabial	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Degrees of constriction	stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	fricative	0	0.02	0.18	0.36	0.39	0.23	0.05	0.07	0.1	0.12	0.18	0.68	0.44	0.54	0.17	0.39	0.22	0.26	0	0
	approximant	0.61	0.61	0.9	0.7	0.8	0.9	0.67	0.65	0.1	0.57	0.93	0.68	0.9	0.9	0.42	0.7	0.7	0.7	0.5	0.5
	high_vowel	0.92	0.9	0.9	0.7	0.8	0.9	1	1	0.93	0.8	0.93	0.69	0.9	0.9	0.9	0.7	0.7	0.7	0.5	0.5
	mid_vowel	0.92	0.9	0.9	0.7	0.8	0.9	1	1	0.93	0.8	0.93	0.69	0.91	0.9	0.9	0.77	0.71	0.7	0.5	0.5
	low_vowel	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 5.3 Results of the measure of pronunciation variation based on the mathematical representation of sound

		Hearer				
		EA	GA	LA	MA	MSA
Speaker	EA		0.127	0.080	0.190	0.160
	GA	0.161		0.101	0.309	0.161
	LA	0.131	0.135		0.333	0.226
	MA	0.208	0.291	0.295		0.362
	MSA	0.154	0.134	0.174	0.335	

Table 5.4 95% confidence intervals for the measure of pronunciation variation based on the mathematical representation of sound

Speaker-Hearer	Degrees of freedom	Mean of normalized distance	Range of 95% confidence interval
EA-GA	206	0.127	0.105 - 0.15
EA-LA	226	0.08	0.064 - 0.096
EA-MA	176	0.19	0.162 - 0.217
EA-MSA	214	0.16	0.131 - 0.19
GA-EA	258	0.161	0.136 - 0.185
GA-LA	288	0.101	0.081 - 0.122
GA-MA	241	0.309	0.275 - 0.343
GA-MSA	282	0.161	0.138 - 0.185
LA-EA	308	0.131	0.112 - 0.15
LA-GA	314	0.135	0.112 - 0.158
LA-MA	266	0.333	0.299 - 0.366
LA-MSA	327	0.226	0.202 - 0.25
MA-EA	183	0.208	0.179 - 0.237
MA-GA	199	0.291	0.254 - 0.327
MA-LA	202	0.295	0.256 - 0.333
MA-MSA	204	0.362	0.322 - 0.401
MSA-EA	205	0.154	0.124 - 0.183
MSA-GA	218	0.134	0.109 - 0.16
MSA-LA	228	0.174	0.147 - 0.2
MSA-MA	188	0.335	0.298 - 0.373

5.3 Measure of Pronunciation variation based on the non-categorical representation of vowels

In this section, the amount of pronunciation variation is determined based on the WPIs calculated following the procedure illustrated in Section 5.1 with the vowels represented by two numbers derived from the formant frequencies as described in Section 2.8. Figure 5.4 shows the distribution of coordinates of the vowels in the first two dimensions of the mathematical representation of sound. The circles show one standard deviation around the mean of the values of the two coordinates representing the vowel categories as calculated in Section 2.8. Solid circles correspond to long vowels and dashed circles correspond to short vowels. The place of articulation of the main articulatory gesture of the vowel is calculated as `(velar + ((palatal-velar)*value derived from F2))`. Similarly, the degree of constriction is calculated as `(high_vowel + ((low_vowel – high_vowel)* value derived from the F1))`. It is important to keep in mind that `velar`, `palatal`, `high_vowel`, and `low_vowel` correspond to parameters of the mathematical representation of sound as discussed in Section 5.1.

Following the algorithm presented in Figure 5.1 with the new representation of vowels, I calculate the WPIs for each pair of varieties. Similar to the procedure in the previous section, I performed the calculation to find the optimal WPIs twice for each pair of varieties. The trials to find the optimal WPIs for all pairs of varieties are provided in Table 5.5. The trials were not carried out in the order given in the table: they are ordered to show the set of WPIs that generated the bigger separation factor and considered the optimal WPIs to show first in the table. The values of `central_vowel` and `mid_vowel` are omitted from Table 5.5 because the new representation is based on calculations that do not include midpoints. The amount of pronunciation variation was calculated based on the algorithm provided in Figure 4.3 with the cost of each basic operation in the Levenshtein distance algorithm based on the optimal WPIs for

the relevant pair of varieties including the non-categorical representation of vowels as discussed earlier. Table 5.6 summarizes the results. Unsurprisingly we see that GA, EA and LA are closer to each other, while MA seems more distant. The same pattern found with all linguistic variation metrics reported in this thesis: the geographically close varieties are also linguistically close. Table 5.7 reports the 95% confidence intervals for the amounts of pronunciation variation reported in Table 5.6. On a par with the findings of the previous measure, EA is closest to MA for both speakers and hearers. There is no distinction regarding the closeness of GA and LA to MA. There is also no distinction regarding the closeness of LA hearers and GA hearers to EA speakers. On the other hand, GA speakers are closer to LA hearers than EA hearers are. Moreover, LA speakers are closer to GA hearers than EA hearers are.

Figure 5.4 Distribution of vowels indicating relevant places of articulation and degrees of constriction to factor the vowels into the mathematical representation of sound

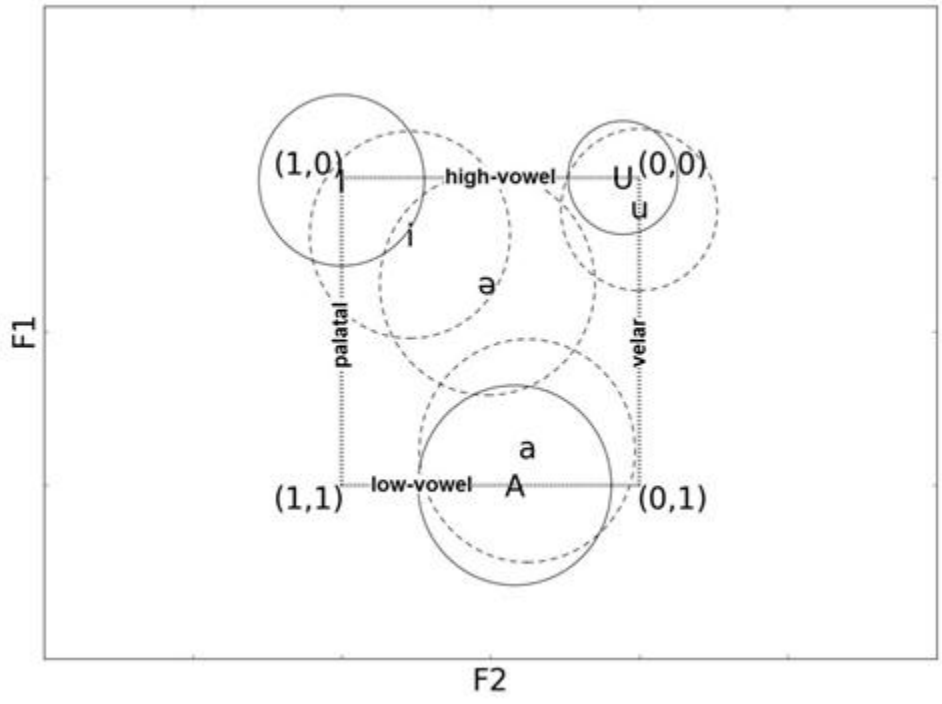


Table 5.5 Summary of the results, based on formants

	pair_name	EA-GA	EA-GA	EA-LA	EA-LA	EA-MA	EA-MA	GA-LA	GA-LA	GA-MA	GA-MA	LA-MA	LA-MA
	Trial number	1	2	1	2	1	2	1	2	1	2	1	2
	separation_factor	2.025	2.024	2.210	2.182	1.843	1.838	2.003	2.002	1.703	1.699	1.796	1.788
	p	0.554	0.768	0.680	0.529	0.632	0.563	0.630	0.643	0.878	0.866	0.950	1.004
	cognates_mean	0.163	0.233	0.181	0.136	0.233	0.202	0.165	0.163	0.350	0.345	0.384	0.410
	cognates_std	0.193	0.265	0.226	0.180	0.217	0.196	0.232	0.240	0.310	0.306	0.315	0.333
	non_cognates_mean	0.915	1.261	1.122	0.850	0.970	0.856	1.106	1.135	1.345	1.326	1.409	1.478
	non_cognates_std	0.178	0.243	0.200	0.147	0.183	0.159	0.237	0.245	0.274	0.271	0.255	0.265
	number of local maxima	168	183	110	39	28	36	122	63	61	45	49	38
Weights	place	3.33	4.96	4.14	2.82	1.99	1.96	5.44	5.78	6.67	6.67	4.78	5
	constriction	1.07	1.23	1.96	1.03	1.1	0.62	0.3	1.11	1.1	1.05	2.77	1.31
	voice	1.47	2.6	2.16	1.51	1.34	1.24	2.2	2.28	1.48	1.5	1.2	1.29
	nasal	1.84	2.6	2.16	1.79	1.69	1.27	2.2	2.28	2.39	2.36	2.52	2.63
	lateral	0.1	0.81	0.71	0.46	0.96	0.71	0.84	0.9	1.52	2.01	0.88	1.89
	trill_flap	1.68	2.6	1.5	1.05	1.33	1.14	1.35	1.5	2.39	2.53	2.68	2.86
	africated	0	0	0	0	0	0.01	0	0	0	0	0	0
	rounded	0.26	0.3	0	0	0.19	0.17	0.51	0.31	1	0.96	0.19	0.29
	long_vowel	0.09	0.18	0	0.06	0	0	0.14	0.19	0.17	0.19	0.08	0.06
	geminated	0.8	0.57	0.03	0.37	1.53	1.53	0	0	0	0	0.01	0
	emphatic	0.4	0.39	0	0	1.69	1.44	0.89	0.87	1.3	1.22	2.26	2.43
	indel	0.84	1.3	1.08	0.77	0.84	0.72	1.1	1.14	1.39	1.36	1.34	1.43
Places of articulation	glottal	0	0	0	0	0	0	0	0	0	0	0	0
	pharyngeal	0.14	0.11	0.16	0.18	0.01	0	0.17	0.15	0.19	0.2	0.03	0.01
	uvular	0.18	0.2	0.4	0.36	0.12	0.08	0.31	0.3	0.4	0.4	0.2	0.2
	velar	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.2	0.2
	central_vowel	-	-	-	-	-	-	-	-	-	-	-	-
	palatal	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.41	0.4	0.48	0.54
	postalveolar	0.5	0.5	0.4	0.4	0.4	0.4	0.67	0.67	0.7	0.7	0.6	0.7
	alveolar	0.5	0.6	0.6	0.6	0.4	0.4	0.7	0.7	0.7	0.7	0.6	0.7
	dental	0.5	0.6	0.6	0.6	0.8	0.8	0.7	0.7	0.7	0.7	0.6	0.7
	labiodental	1	0.9	1	1	1	1	1	1	1	1	0.93	0.97
	bilabial	1	1	1	1	1	1	1	1	1	1	1	1
Degrees of constriction	stop	0	0	0	0	0	0	0	0	0	0	0	0
	fricative	0.01	0.1	0.29	0.4	0.29	0.32	0.19	0.08	0.12	0.15	0.29	0.71
	approximant	0.89	0.6	0.57	0.88	0.9	0.6	0.19	0.08	0.8	0.67	0.53	0.8
	high_vowel	0.9	0.8	0.73	0.9	0.9	0.9	0.31	0.9	0.8	0.8	0.9	0.8
	mid_vowel	-	-	-	-	-	-	-	-	-	-	-	-
	low_vowel	1	1	1	1	1	1	1	1	1	1	1	1

Table 5.6 Results of the measure of pronunciation variation based on the non-categorical representation of vowels

		Hearer				
		EA	GA	LA	MA	
Speaker	EA		0.009	0.009	0.017	
	GA	0.011		0.007	0.024	
	LA	0.012	0.009		0.026	
	MA	0.018	0.023	0.024		
	(MSA excluded due to lack of acoustic data.)					

Table 5.7 95% confidence intervals for the measure of pronunciation variation based on the non-categorical representation of sound

Speaker-Hearer	Degrees of freedom	Mean of normalized distance	Range of 95% confidence interval
EA-GA	377	0.009	0.008 - 0.011
EA-LA	414	0.009	0.008 - 0.011
EA-MA	323	0.017	0.015 - 0.019
GA-EA	356	0.011	0.01 - 0.013
GA-LA	396	0.007	0.006 - 0.008
GA-MA	332	0.024	0.021 - 0.026
LA-EA	406	0.012	0.01 - 0.013
LA-GA	412	0.009	0.007 - 0.01
LA-MA	349	0.026	0.024 - 0.028
MA-EA	322	0.018	0.016 - 0.02
MA-GA	348	0.023	0.021 - 0.025
MA-LA	353	0.024	0.021 - 0.026

CHAPTER 6

CONCLUSION

In this thesis, I proposed a new methodology to computationally measure the amount of lexical and pronunciation variation between five varieties of Arabic. I argued for measuring the amount of linguistic variation asymmetrically. I used two tests of reliability with a convergence test and statistical tools. I also developed a new representation of sound for measuring pronunciation similarity that is mathematically based and computationally effective. This representation is able to represent sound categorically and non-categorically. Moreover, it has the ability to dynamically reflect the patterns of sound change based on pronunciation similarity of cognate words and pronunciation dissimilarity of non-cognate words. I incorporated the new representation of sound in two measures of pronunciation variation using the Levenshtein distance algorithm. I also implemented an optimization technique to set the costs of insertions, deletions, and substitutions of the Levenshtein distance algorithm, with the cost of substitution derived from the mathematical representation of sound. This allows the cost to be dynamically calculated based on the pronunciation similarity of the sounds being substituted.

I developed two computational measures of lexical variation and three computational measures of pronunciation variation based on native speaker elicitations of the Swadesh list. The first computational measure of lexical variation was based on whether the hearer's variety has a cognate of the speaker variety's words for the same Swadesh list item. The second measure of lexical variation incorporated the pronunciation variation of the words by comparing their transcriptions in IPA. The first measure of pronunciation variation was phonemic where the costs of the basic operations of the Levenshtein distance algorithm were set to a default cost. The

second measure of pronunciation variation took into account phonetic similarity by incorporating the mathematical representation of sound in the calculation of the basic operations. The third measure of pronunciation variation used a non-categorical representation of vowels derived from the values of the first and second formant frequencies.

All measures of linguistic variation developed in this thesis showed a consistent pattern where the geographically closer varieties tend to be also linguistically closer: EA, LA, and GA tend to be closer to each other than to MA. We also consistently found two patterns of asymmetry in the results of the lexical and pronunciation variation metrics. The asymmetry is indicated when the amount of variation between a speaker of a variety X and a hearer of a variety Y is not equal to the amount of variation between a speaker of a variety Y and a hearer of a variety X. The first pattern shows that the amounts of variation between EA speakers and hearers of GA, LA, and MA are less than the amounts of variation between EA hearers and speakers from the corresponding varieties. This reflects a pattern of mutual intelligibility we observe in the communication of Egyptians with members of other local varieties. The Egyptian speakers are understood better than they understand other speakers. This leads speakers of other varieties to accommodate Egyptian speakers in most cases. The second pattern of asymmetry shows that the amounts of variation between LA hearers and speakers of EA, GA, and MA are less than the amounts of variation between LA speakers and hearers from the corresponding varieties. This could imply that members of the LA variety are able to understand members of other varieties better than the other varieties understand them. The two claims regarding the patterns of asymmetry of the variation metrics for EA and LA speakers require verification by an independent study of mutual intelligibility.

The results of the first measure of lexical variation show that the closest variety to MSA is LA followed by both GA and EA. On the other hand, GA is measured to be the closest to MSA based on the other variation metrics that considered MSA. All variation metrics have indicated that MA is the farthest to MSA (see Section 6.1 for relevant discussion). The lexical and pronunciation variation metrics at the phonemic level resulted with LA second closest and EA third. The pronunciation variation metric at the phonetic level resulted with EA second and LA third. The first row in Table 6.1 shows the order of the closeness of the local varieties to MSA. The distinction in the measurement is considered not significant if there is an overlap in the 95% confidence intervals for the measurements. The non-significance in the difference of the closeness to MSA is indicated by grouping the varieties between braces or parentheses. For example, '{GA, (LA}, EA), MA' means that the closest to MSA is GA followed by LA, EA, then MA. However, there is an overlap in the confidence intervals for '{GA, LA}' and there is an overlap in the confidence intervals for '(LA, GA)'.

The second and third rows in Table 6.1 show the amount of variation between MA and the other local varieties. None of the variation metrics provided results that distinguish the closeness of GA and LA to MA. Therefore, GA and LA are grouped between braces. The second row indicates whether EA speakers are closer to MA hearers than GA and LA speakers to MA hearers. The third row indicates whether MA speakers are closer to EA hearers than MA speakers to GA and LA hearers. The lexical variation metric at the phonemic level did not provide any distinction regarding the closeness of the local varieties to MA. As can be seen in the second and third rows in Table 6.1, most pronunciation variation metrics developed in this research have indicated that EA is closer to MA than both GA and LA in both directions of communication. The fourth row shows that all variation metrics but the pronunciation variation

with non-categorical representation of vowels have indicated that EA speakers are closer to LA hearers than GA hearers are. The fifth row shows that, according to all variation metrics, GA speakers are closer to LA hearers than EA hearers are. Finally, only according to the pronunciation variation metric with the non-categorical representation of vowels, LA speakers are significantly closer to GA hearers than EA hearers are, as the sixth row shows.

Table 6.1 Summary of the closeness of the Arabic varieties to each other

Row number	The comparison key	Lexical at the phonemic level	Pronunciation at the phonemic level	Pronunciation at the phoneic level	Pronunciation with non-categorical vowel representation
1	Order of closeness to MSA	{GA, (LA}, EA), MA	GA, {LA, EA}, MA	{GA, EA, LA}, MA	No MSA data
2	EA-MA < {GA, LA}-MA	Not Significant	YES	YES	YES
3	MA-EA < MA-{GA, LA}	Not Significant	Not Significant	YES	YES
4	EA-LA < EA-GA	YES	YES	YES	Not Significant
5	GA-LA < GA-EA	YES	YES	YES	YES
6	LA-GA < LA-EA	Not Significant	Not Significant	Not Significant	YES

6.1 The limited representation of the Arabic varieties

As mentioned in Chapter 2, each local variety is represented by only two male native speakers from a major city where the variety under consideration is spoken. Other speakers from the same city or from other cities have different lexical inventories and different pronunciations to some degree. Moreover, the amount of variation is expected to show different patterns if rural areas are considered. The number of speakers and the geographical representation is considered a limitation of the study; these results would not necessarily generalize to other areas where the varieties are spoken. Also, the representation of MSA is derived from two modern dictionaries of Arabic, which does not necessarily capture all possible translations of the words of the Swadesh list. Moreover, the two modern dictionaries were authored by LA speakers, which raises the question of whether that biased MSA to be closer the LA to some degree? I selected those two dictionaries after careful consideration of the quality of their translations, with any bias expected to be marginal. However, it would still be worthwhile to see the effect on the results using dictionaries developed by speakers of other varieties.

6.2 Implications of different local maxima

The optimization technique of the separation factor we followed produced a large number of maxima for each pair of varieties. The set of WPIs that generates the largest separation factor was selected as the optimal set of WPIs. However, other local maxima were not too remote from the selected optimal maximum – they also provided a meaningful representation of WPIs. Different local maxima can be seen as competing in identifying the right cost for different sets of combinations of sound changes. The large number of factors including the large number of pairs of varieties, the large number of pairs of words, and the large number of identified local maxima

makes it impossible to present all combinations of results in this thesis. I will focus on one pair of varieties and present a subset of the findings related to it; similar patterns are found for other pairings. The attempt to identify the optimal WPIs for the pair of varieties (LA and GA) resulted with 94 local maxima with the value of the separation factor ranging from 1.81 to 1.97, all of which are reasonable approximations. The total number of pairs of words included here is 1155, consisting of 398 non-cognate words and 757 cognate words. 19 out of the 757 pairs of cognate words were not identified correctly by any of the 94 local maxima. On the other hand, 663 pairs of cognate words were identified as cognates correctly by all local maxima. The focus of the following discussion is on a sample drawn from the 75 pairs of cognate words that were identified correctly as cognates by a subset of the local maxima.

Table 6.2 contains a sample of cognate words divided into two categories based on the sound changes taking place in them. Category 1 shows a sample of the pairs of words that include an assimilation of ʔ to *w* or *g*. Pairs in this category were identified correctly by only 4 local maxima, but those local maxima failed to correctly identify cases of pairs reported in Table 6.2 under Category 2. In this case, the optimization technique failed to find a set of WPIs that could identify all words in Table 6.2; it could be the case that such set of WPIs does not exist.

Table 6.2 Sample from LA-GA data set

Levantine	Gulf	Category	Number of local maxima able to identify
ʔIʃ	wIʃ	1	4
ʔAl	gAl	1	4
tallaʒ	θalladʒ	2	90
taliʒ	θaldʒ	2	65

6.3 Computational limitations

The computational complexity of the algorithm used to find the optimal WPIs dictated some constraints to achieve a computationally feasible solution, and potential alternatives should be considered in the future. Increasing the scope from which the starting random values are selected could result in different optimal WPIs. Having a smaller step size to begin with could change the results as well. Increasing the number at which we stop the process of finding more local maxima might also result in a more accurate model.

It is also possible to add more dimensions by specifying a more fine-grained cost of indels. The fundamental difference between insertions and deletions requires a thoughtful review of the matter. In deletions, the hearer is missing information that needs to be recovered, whereas in insertions the hearer is getting extra information that needs to be deleted. Consider for example the verbs *rama* and *rma*, meaning 'threw' in LA and MA respectively. A speaker of MA produces the verb missing a vowel. In such a case, the LA hearer has to figure out the missing vowel and recover it. In the opposite direction of communication, a speaker of LA produces the verb with one extra vowel according to the MA. The assumption is that the MA hearer will have less difficulty deleting the extra information – the second vowel in this case – than the LA hearer who has to recover the missing information. This fundamental difference suggests a higher cost for deletions than insertions. In addition, inserting or deleting a vowel does not necessarily need to be equal to the cost of inserting or deleting a consonant, even though confidently setting a specific cost relative to the phonological operation (insertion or deletion) or category (consonant or vowel) may be open to debate. Assigning the same cost for both operations (insertion and deletion) has been the norm in previous research, as well as having a symmetric variation metric. However, in Heeringa and Braun (2003) the cost of insertion of a

vowel is assumed to be the cost of replacing that vowel with a schwa with the vowel feature set to 0. For consonants, they used a glottal stop with the consonant feature set to 0. From the perspective of the present study, this problem could be resolved by adding more dimensions to calculate more fine-grained cost of indels. For example, we could have four dimensions to evaluate – independent of each other – the costs of inserting a vowel, deleting a vowel, inserting a consonant, and deleting a consonant. Even a more fine-grained and computationally very expensive solution is to have a distinct dimension for inserting each phoneme and a distinct dimension for deleting each phoneme. Such solution is not feasible given the available computational resources.

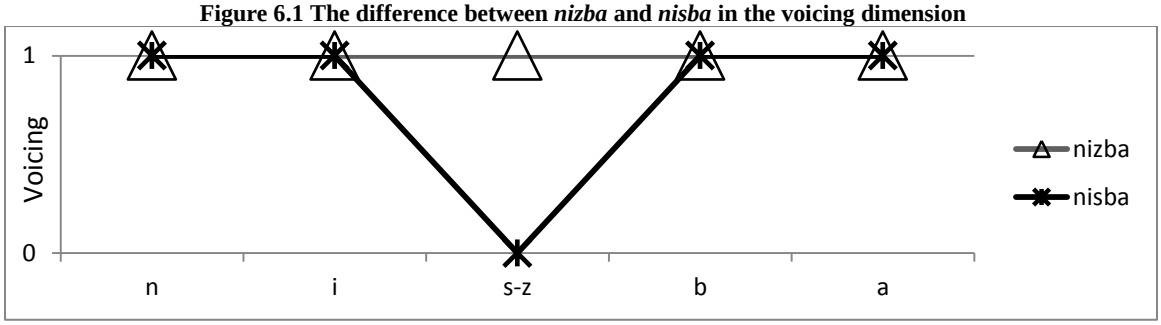
The mathematical representation of sound represents each phonetic feature and articulatory gesture by a distinct dimension in a multidimensional space. Geometrically speaking, each dimension is perpendicular on all other dimensions. This implies that the phonetic features and articulatory gestures are considered independent where a change in one of them does not carry any effect on the rest. This assumption is not motivated by linguistic theory; rather it is made for the sake of computational simplicity. Some features relate to others, such as rounding and backness in vowels. This imposes a limitation on the current study because these features are assumed to be independent, whereas a more accurate model would consider these interdependencies. Certainly, this is an area to be explored in further research.

6.4 Patterns of sound change and the mathematical representation of sound

One of the most important factors of sound change is phonetic feature overlap or articulatory gesture overlap. An example of phonetic feature overlap is the assimilation of voicing when a voiceless fricative occurs between voiced segments. This is the case for example,

with the cognate words *nisba* and *nizba* ‘percentage’ in LA and EA respectively. The voicing feature is introduced in the third phoneme of the Egyptian word because it occurs in the context of voiced phonetic segments. The two paths representing the words in the multidimensional representation are almost identical except for the third phoneme where there is a shift in the voicing dimension. Focusing on the dimension representing the voicing feature (Figure 6.1) we see that the path of the second word is always set to one value marking that all the phonemes are voiced, while in the first word the third phoneme is set to a voiceless value. Such sound change can be computationally approximated as smoothing the path connecting the points representing the phonemes, specifically in the voicing dimension. Some examples of sound change that can be accounted for as smoothing of the line connecting the pointes representing the phones in the mathematical representation of sound are the spread of the emphatic feature in the dialects of Arabic, vowel nasalization in context of nasals in American English, velarized nasals in the context of velars, and many other examples. It could be argued that such sound change when it is phonologically derived by features from neighboring segments should have a different penalty because it is considered as a natural sound change. This is certainly an important topic that deserves further investigations.

Figure 6.1 The difference between *nizba* and *nisba* in the voicing dimension



6.5 Suggestions for future research

Based on the results of this research, I suggest that future research should continue to investigate new representations of acoustic segments. Mainly, representations that encode phonetic features and/or articulatory gestures. From an abstraction point of view, such representations are more detailed than the phonemic representation and more abstract – and less complex than – than the acoustic representation. I implemented a mathematical representation based on a hybrid model derived from articulatory phonology and autosegmental phonology. More significantly, I developed a model that ranks the fitness of the representation of sound based on its ability to reliably identify cognate words. This model can be used to investigate the fitness of other representations of sound based on other, or combinations of other, phonological theories. In addition, there is a room for improving the optimization technique by using more sophisticated computational methods.

Because the focus of the current project is to measure the linguistic variation between a set of Arabic varieties, the model was kept as simple as possible to accomplish the task at hand. That said, the mathematical representation of sound could easily be enhanced to accommodate more complicated sound representations. The different degrees in which the phonetic features are manifested can be encoded in a multivalued scale in their corresponding dimension. For example, different types of phonation, such as creaky voicing, can be distinguished by providing more than one value in the voicing dimension. Another enhancement is to represent an utterance as a line in a multi-dimensional space. The model presented in this thesis represents utterances as sequence of points, whereas connecting the points in a way that reflects the transition of the phonetic features and articulatory gestures between the phonemes could result in a model that is more generally applicable. For the purposes of measuring the amount of variation, such a model

has consequences for the way distance is measured. Of course, here are consequences for each adaptation of the model, and these are left for future research.

Another promising direction for extending this methodology would be to add an alignment function to the algorithm. This would be useful to identifying correspondences in the cognate words that would generalize to sound changes in the dialects. It is possible this could eventually be extended to represent historical relationships and even help with historical reconstruction. At the very least, it could provide useful metrics for comparing the differences between varieties which could in turn be helpful to pedagogical and computational approaches to language variation. Being able to automatically adjust speech recognition systems trained on one variety to recognize another could have wider application potential.

REFERENCES

Abunasser, M., Mahrt, T., & Shih, C. (2011, May). *Arabic Emphatics: What happens with lips and tongue*. Poster presented at the Speech Production Workshop. University of Illinois at Urbana-Champaign, Urbana, IL.

Almeida, A., & Braun, A. (1986). "Richtig" und"Falsch" in Phonetischer Transkription. Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, 158-172.

Ba'albaki, R., & Ba'albaki, M. (1999). *Al-Mawrid*. Beirut, Lebanon: Dar El-ilm Lilmalayin.

Babitch, R. M., & Lebrun, E. (1989). Dialectometry as computerized agglomerative hierarchical classification analysis. *Journal of English Linguistics*, *22*(1), 83-87.

Berghel, H., & Roach, D. (1996). An extension of Ukkonen's enhanced dynamic programming ASM algorithm. *ACM Transactions on Information Systems (TOIS), 14(1)*, 94-106.

Biadsy, F., Hirschberg, J., & Habash, N. (2009, March). Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages* (pp. 53-61). Association for Computational Linguistics.

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer* [Computer program]. Version 5.3.35, retrieved 8 December 2012 from http://www.praat.org/

Cadora, F. J. (1979). *Interdialectal lexical compatibility in Arabic: an analytical study of the lexical relationships among the major Syro-Lebanese varieties* (Vol. 11). Brill Archive.

Covington M. A. (1996). An algorithm to align words for historical comparison. *Computational Linguistics, 22(4),* 481-496

Cucchiarini, C. (1983). *Phonetic transcription: A methodological and empirical study*. PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods, 41(2),* 385-390.

Ebobisse, C. (1989). Dialectométrie lexicale des parlers sawabantu. *Journal of West African Languages, 19*, 2:57-66.

Elfardy, H., & Diab, M. (2013, August). Sentence Level Dialect Identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 456-461). Association for Computational Linguistics.

Elias, E. A., & Elias, E. E. (1983). *Elias' Modern Dictionary, Arabic-English.* Cairo, Egypt: Elias Modern Publishing House.

Elsie, R. (1986). *Dialect Relationships in Goidelic: A Study in Celtic Dialectology*. Helmut Buske, Hamburg.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science, 155(760),* 279-284.

Gildea, D., & Jurafsky, D. (1996). Learning bias and phonological-rule induction. *Computational Linguistics*, *22*(4), 497-530.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and multicultural development, 28(6),* 445-467.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426(6965),* 435-439.

Gray, R. D., & Jordan, F. M. (2000). Language trees support the express-train sequence of Austronesian expansion. *Nature, 405(6790),* 1052-1055.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica, 45*(1), 31-80.

Heeringa, W. J. (2004). *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Doctoral dissertation, University Library Groningen.

Heeringa, W., & Braun, A. (2003). The use of the Almeida-Braun system in the measurement of Dutch dialect distances. *Computers and the Humanities*,*37*(3), 257-271.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Hoppenbrouwers, C. A. J., & Hoppenbrouwers, G. A. (2001). *De indeling van de Nederlandse streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Uitgeverij Van Gorcum.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics* (pp. 60-66). Morgan Kaufmann Publishers Inc..

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, *37*(3), 273-291.ucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study*.

Kondrak, G. (2009). Identification of cognates and recurrent sound correspondences in word lists. *Traitement automatique des langues*, *50*(2), 201-235.

Kondrak, G., & Sherif, T. (2006, July). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of the Workshop on Linguistic Distances* (pp. 43-50). Association for Computational Linguistics.

Ladefoged P. (1975). *A course in Phonetics.* Harcourt Brace Jovanovich, New York.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, *10(8):* 707-710.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, *58(4),* 880-883.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33(1)*, 31-88.

Nearey, T. M. (1978). *Phonetic feature systems for vowels* (Vol. 77). Indiana University Linguistics Club.

Nerbonne J., Heeringa W. (1997) Measuring Dialect Distance Phonetically. In *Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology*.

Nerbonne, J., & Kretzschmar, W. (2003). Introducing computational techniques in dialectometry. *Computers and the Humanities*, *37(3),* 245-255.

Oakes M. P. (2000) Computer Estimation of Vocabulary in Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics, 7(3),* pp. 233-243.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, *37*:1–24.

Serva, M., & Petroni, F. (2008). Indo-European languages tree by Levenshtein distance. *EPL (Europhysics Letters)*, *81*(6), 68005.

Somers H. L. (1998) Similarity Metrics for Aligning Children's Articulation Data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 1227-1231.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Thomas, E R. & Kendall, T. (2007). *NORM: The vowel normalization and plotting suite*. [Online Resource: http://ncslaap.lib.ncsu.edu/tools/norm/]

Ukkonen, E. (1983). On approximate string matching. In *Foundations of Computation Theory* (pp. 487-495). Springer Berlin/Heidelberg.

Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and control, 64(1),* 100-118.

Valls, E., Nerbonne, J., Prokic, J., Wieling, M., Clua, E., & Lloret, M. R. (2011). Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches. *Verba: Anuario Galego de Filoloxía, 39.*

Vieregge, W. H., Rietveld, A. C., & Jansen, C. (1984). A distinctive feature based system for the evaluation of segmental transcription in Dutch. In *Proc. of the 10th International Congress of Phonetic Sciences* (pp. 654-659). Foris Publications.

Wagner, H. (1958). *Linguistic atlas and survey of Irish dialects*. Dublin Institute for advanced studies.

Wichmann, S., Holman, E. W., Bakker, D., & Brown, C. H. (2010). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, *389*(17), 3632-3639.

Zaidan, O. F., & Callison-Burch, C. (2012). Arabic dialect identification. *Computational Linguistics. 52(1).* Association for Computational Linguistics.

APPENDIX A

THE SWADESH LIST FOR THE VARIETIES OF ARABIC UNDER CONSIDERATION

This appendix lists the elicited lexical items for all participants in this study. The list consists of 207 items of the Swadesh list. Each item is given an ID, and context sentence along with the English word. Those are listed in the first column titled "Swadesh item info". The remaining columns correspond to:

- Speaker ID: each participant is given an ID that consists of 4 characters. The first two characters correspond to the variety abbreviation and the second two characters are a sequence number 01 or 02.
- Word number: a sequence number of the translations provided for each lexical item by each participant.
- Source of stimulus: an ID (ENG or VAR). ENG identifies the source of stimulus as the English word and context sentence given in the first pass. VAR identifies the source of stimulus as the word provided by other participants along with the English word and context sentence, elicited in the second pass.
- Word in Arabic script: The transcription of the words in Arabic script according to the guidelines described in section 2.6.
- Word in IPA script: The IPA transcription of the words. Long vowels are encoded by upper case letters and gemination feature for consonants is encoded by '+' sign.

Word origin ID: a unique ID for each set of cognate words under each Swadesh list item. This ID is assigned to each elicited item based on the researcher's knowledge of the language.

Swadesh item info	Speaker ID	Word Number	Source of stimulus	Word in Arabic script	Word in IPA	Word origin ID
ID: SWADESH_001						
English word: I						
Context: ___ like the book						
	EA01	1	ENG	عَنا	ʔana	1
	EA02	1	ENG	عَنا	ʔana	1
	GA01	1	ENG	عَنا	ʔana	1
	GA02	1	ENG	عَنا	ʔana	1
	LA01	1	ENG	عَنا	ʔana	1
	LA02	1	ENG	عَنا	ʔana	1
	MA01	1	ENG	عَنا	ʔana	1
	MA02	1	ENG	عَنا	ʔana	1
	SA01	1	ENG	عَنا	ʔanA	1
ID: SWADESH_002						
English word: you						
Context: (Talking to Ali.) ___ like the book						
	EA01	1	ENG	عِنتا	ʔinta	1
	EA02	1	ENG	عِنتا	ʔinta	1
	GA01	1	ENG	عَنتا	ʔanta	1
	GA02	1	ENG	عِنتا	ʔinta	1
	LA01	1	ENG	عِنتا	ʔinta	1
	LA02	1	ENG	نتا	nta	1
	MA01	1	ENG	نتا	nta	1
	MA02	1	ENG	نتا	nta	1
	SA01	1	ENG	عَنتَ	ʔanta	1
ID: SWADESH_003						
English word: he						
Context: ___ likes the book						
	EA01	1	ENG	هُوَّ	huw+a	1
	EA02	1	ENG	هُوَّ	huw+a	1

GA01	1	ENG	هُوّا	huw+a	1
GA02	1	ENG	هو	hu	1
LA01	1	ENG	هُوَّ	huw+a	1
LA02	1	ENG	هُوُّ	huw+ə	1
MA01	1	ENG	هُوّا	huw+a	1
MA02	1	ENG	هُوّا	huw+a	1
SA01	1	ENG	هُوَ	huwa	1

ID: SWADESH_004
English word: we
Context: ___ like the book

EA01	1	ENG	ءِحنا	ʔiħna	1
EA02	1	ENG	ءِحنا	ʔiħna	1
GA01	1	ENG	حْنّا	ħən+a	1
GA02	1	ENG	ءِحنا	ʔiħna	1
LA01	1	ENG	ءِحنا	ʔiħna	1
LA02	1	ENG	نِحنا	niħna	1
MA01	1	ENG	حنا	ħna	1
MA02	1	ENG	حنا	ħna	1
SA01	1	ENG	نَحنُ	naħnu	1

ID: SWADESH_005
English word: you
Context: (Talking to a group of 5 boys.) ___ like the book

EA01	1	ENG	ءِنتُم	ʔintum	1
EA01	2	VAR	ءِنتو	ʔintu	1
EA02	1	ENG	ءِنتو	ʔintu	1
GA01	1	ENG	ءَنتُم	ʔantum	1
GA02	1	ENG	ءِنتون	ʔintUn	1
LA01	1	ENG	ءِنتو	ʔintu	1
LA02	1	ENG	ءنتو	ʔntu	1
MA01	1	ENG	نتوما	ntUma	1
MA02	1	ENG	نتوما	ntUma	1
SA01	1	ENG	ءَنتُم	ʔantum	1

ID: SWADESH_006
English word: they
Context: (Talking about a group of 5 boys.) ___ like the book

EA01	1	ENG	هُمّا	hum+a	1
EA02	1	ENG	هُمّا	hum+a	1
GA01	1	ENG	هُم	hum	1
GA02	1	ENG	هُم	hum	1
LA01	1	ENG	هُمّا	hum+a	1

LA02	1	ENG	هِنّْ	hin+ə	1
LA02	2	ENG	هِنّْن	hin+ən	1
MA01	1	ENG	هوما	hUma	1
MA02	1	ENG	هوما	hUma	1
SA01	1	ENG	هُم	hum	1

ID: SWADESH_007
English word: this
Context: (Pointing) ___ is a book

EA01	1	ENG	دَ	da	1
EA02	1	ENG	دَ	da	1
GA01	1	ENG	هاذا	hAða	1
GA02	1	ENG	دَ	da	1
LA01	1	ENG	هاظ	hAðˁ	1
LA02	1	ENG	هَيدا	hayda	1
MA01	1	ENG	هادا	hAda	1
MA02	1	ENG	هادا	hAda	1
SA01	1	ENG	هاذا	hAðA	1

ID: SWADESH_008
English word: that
Context: (Pointing) ___ is a book

EA01	1	ENG	دَ	da	1
EA02	1	ENG	دَ	da	1
GA01	1	ENG	هاذاك	hAðAk	1
GA02	1	ENG	دَ	da	1
LA01	1	ENG	هَظاك	haðˁAk	1
LA02	1	ENG	هَيداك	haydAk	1
MA01	1	ENG	هادَك	hAdak	1
MA02	1	ENG	هادَك	hAdak	1
SA01	1	ENG	ذاكَ	ðAka	1

ID: SWADESH_009
English word: here
Context: (Pointing) ___, on the table exists a book

EA01	1	ENG	هِنا	hina	1
EA02	1	ENG	هِنا	hina	1
GA01	1	ENG	هِنا	hina	1
GA02	1	ENG	عَكّو	ʔak+u	2
LA01	1	ENG	هون	hUn	1
LA01	2	ENG	هونَ	hUna	1
LA02	1	ENG	هون	hUn	1
MA01	1	ENG	هنا	hna	1

MA02	1	ENG	هنا	hna	1
SA01	1	ENG	هُنا	hunA	1

ID: SWADESH_010
English word: there
Context: (Pointing) ___, in the room exists a book

EA01	1	ENG	هِناك	hinAk	1
EA02	1	ENG	هِناك	hinAk	1
GA01	1	ENG	هِناك	hinAk	1
GA02	1	ENG	عَكّو	ʔak+u	2
LA01	1	ENG	هُناك	hunAk	1
LA02	1	ENG	هونيك	hUnIk	1
MA01	1	ENG	تَمّا	tam+a	3
MA01	2	VAR	هناك	hnAk	1
MA02	1	ENG	تَمّا	tam+a	3
MA02	2	ENG	لهيه	lhIh	4
SA01	1	ENG	هُناكَ	hunAka	1
SA01	2	ENG	ثَمَّ	θam+a	3

ID: SWADESH_011
English word: who
Context: ___ closed the door?

EA01	1	ENG	مين	mIn	1
EA02	1	ENG	مين	mIn	1
GA01	1	ENG	مَن	man	1
GA02	1	ENG	مْن	mən	1
LA01	1	ENG	مين	mIn	1
LA01	2	VAR	مَنو	manU	1
LA02	1	ENG	مين	mIn	1
MA01	1	ENG	شكون	ʃkUn	2
MA02	1	ENG	شكون	ʃkUn	2
SA01	1	ENG	مَن	man	1

ID: SWADESH_012
English word: what
Context: ___ did you eat yesterday?

EA01	1	ENG	ءيه	ʔI	1
EA02	1	ENG	ءيه	ʔIh	1
GA01	1	ENG	ءيش	ʔIʃ	1
GA01	2	VAR	شِنهو	ʃinhu	1
GA02	1	ENG	ويش	wIʃ	1
GA02	2	VAR	شُنو	ʃunu	1
LA01	1	ENG	شو	ʃu	1

LA01	2	ENG	ءيش	ʔIʃ	1
LA02	1	ENG	شو	ʃu	1
MA01	1	ENG	شنو	ʃnu	1
MA02	1	ENG	شنو	ʃnu	1
SA01	1	ENG	ماذا	mAðA	2

ID: SWADESH_013
English word: where
Context: ___ did you go yesterday?

EA01	1	ENG	فين	fIn	1
EA02	1	ENG	فين	fIn	1
GA01	1	ENG	وين	wIn	1
GA02	1	ENG	وين	wIn	1
LA01	1	ENG	وين	wIn	1
LA02	1	ENG	وين	wIn	1
MA01	1	ENG	فين	fIn	1
MA02	1	ENG	فين	fIn	1
SA01	1	ENG	عَينَ	ʔayna	1

ID: SWADESH_014
English word: when
Context: ___ did you eat your breakfast?

EA01	1	ENG	ءِمتا	ʔimta	1
EA02	1	ENG	ءِمتا	ʔimta	1
GA01	1	ENG	مَتا	mata	1
GA02	1	ENG	مَتا	mata	1
LA01	1	ENG	مَتا	mata	1
LA01	2	ENG	ءِمتا	ʔimta	1
LA02	1	ENG	ءِمتا	ʔImta	1
MA01	1	ENG	فوقاش	fUqAʃ	2
MA02	1	ENG	ءِمتا	ʔimta	1
MA02	2	VAR	وَقتاش	waqtAʃ	2
SA01	1	ENG	مَتا	matA	1

ID: SWADESH_015
English word: how
Context: ___ did you repair the car?

EA01	1	ENG	ءِزّاي	ʔiz+Ay	1
EA02	1	ENG	ءِزّاي	ʔiz+Ay	1
GA01	1	ENG	كيف	kIf	2
GA01	2	VAR	شلون	ʃlUn	3
GA02	1	ENG	شلون	ʃlUn	3
LA01	1	ENG	كيف	kIf	2

LA01	2	VAR	شلون	ʃlUn	3
LA02	1	ENG	كيف	kIf	2
MA01	1	ENG	كيفاش	kIfAʃ	2
MA02	1	ENG	كيفاش	kIfAʃ	2
SA01	1	ENG	كَيفَ	kayfa	2

ID: SWADESH_016
English word: not
Context: Negating an adjective as in: The boy is ___ tall

EA01	1	ENG	مِش	miʃ	1
EA02	1	ENG	مِش	miʃ	1
GA01	1	ENG	ماهو	mAhu	2
GA01	2	ENG	مُهُب	məhub	2
GA01	3	VAR	مُش	məʃ	1
GA02	1	ENG	مو	mU	2
LA01	1	ENG	مِش	miʃ	1
LA02	1	ENG	مَنّو	man+u	2
LA02	2	VAR	مِش	miʃ	1
MA01	1	ENG	ماشي	mAʃi	1
MA02	1	ENG	ماشي	mAʃi	1
SA01	1	ENG	لَيسَ	laysa	3

ID: SWADESH_017
English word:
all
Context: ___ boys are tall

EA01	1	ENG	كُل	kul	1
EA02	1	ENG	كُل	kul	1
GA01	1	ENG	كِل	kil	1
GA02	1	ENG	كِل	kil	1
LA01	1	ENG	كُل	kul	1
LA02	1	ENG	كِل	kil	1
MA01	1	ENG	گاع	gAʕ	2
MA02	1	ENG	گاع	gAʕ	2
SA01	1	ENG	كُلّ	kul+	1

ID: SWADESH_018
English word: many
Context: ___ boys are tall

EA01	1	ENG	كُتير	kətIr	1
EA02	1	ENG	كُتير	kətIr	1
GA01	1	ENG	كثير	kθIr	1
GA01	2	VAR	واجِد	wAdʒid	2

GA02	1	ENG	واحِد	wAdʒid	2
LA01	1	ENG	كثير	kθIr	1
LA02	1	ENG	كتير	ktIr	1
MA01	1	ENG	بْزّاف	bəz+Af	3
MA02	1	ENG	بْزّاف	bəz+Af	3
SA01	1	ENG	كَثير	kaθIr	1

ID: SWADESH_019
English word: some
Context: ___ boys are tall

EA01	1	ENG	شوَية	ʃwaya	1
EA02	1	ENG	شْوَية	ʃəwaya	1
GA01	1	ENG	بَعظ	baʕðˁ	2
GA02	1	ENG	بَعظ	baʕðˁ	2
LA01	1	ENG	بَعظ	baʕðˁ	2
LA01	2	ENG	شوَيَّة	ʃway+ə	1
LA02	1	ENG	بَعض	baʕdˁ	2
LA02	2	VAR	شوَيِّة	ʃway+ih	1
MA01	1	ENG	شوَيَّة	ʃway+a	1
MA01	2	VAR	بَعض	baʕdˁ	2
MA02	1	ENG	شي	ʃi	1
MA02	2	VAR	بَعض	baʕdˁ	2
SA01	1	ENG	بَعض	baʕdˁ	2

ID: SWADESH_020
English word: few
Context: ___ boys are tall

EA01	1	ENG	كَم	kam	1
EA01	2	VAR	بَعض	baʕdˁ	2
EA01	3	VAR	شْوَية	ʃəwaya	3
EA02	1	ENG	كَم	kam	1
GA01	1	ENG	گْليل	gəlll	4
GA01	2	ENG	شوَي	ʃway	3
GA02	1	ENG	بَعظ	baʕðˁ	2
LA01	1	ENG	شوَيَّة	ʃway+ə	3
LA01	2	ENG	گليل	galll	4
LA02	1	ENG	شوَيِّة	ʃway+ih	3
LA02	2	VAR	كَم	kam	1
MA01	1	ENG	شوَيَّة	ʃway+a	3
MA02	1	ENG	قليل	qlll	4
MA02	2	ENG	شي	ʃi	3
SA01	1	ENG	بَعض	baʕdˁ	2
SA01	2	ENG	قليل	qlll	4

ID: SWADESH_021
English word: other
Context: I am not looking for this book,I am looking for the ___ book

EA01	1	ENG	تاني	tAni	1
EA02	1	ENG	تاني	tAni	1
GA01	1	ENG	ثاني	θAni	1
GA02	1	ENG	ثاني	θAni	1
LA01	1	ENG	ثاني	θAni	1
LA02	1	ENG	تاني	tAni	1
MA01	1	ENG	ءاخُر	ʔAxur	2
MA02	1	ENG	ءاخُر	ʔAxur	2
SA01	1	ENG	ءاخَر	ʔAxar	2

ID: SWADESH_022
English word: one
Context: This is number ___

EA01	1	ENG	واحِد	wAħid	1
EA02	1	ENG	واحد	wAħd	1
GA01	1	ENG	واحِد	wAħid	1
GA02	1	ENG	واحِد	wAħid	1
LA01	1	ENG	واحَد	wAħad	1
LA02	1	ENG	واحَد	wAħad	1
MA01	1	ENG	واحْد	wAħəd	1
MA02	1	ENG	واحد	wAħd	1
SA01	1	ENG	واحِد	wAħid	1

ID: SWADESH_023
English word: two
Context: This is number ___

EA01	1	ENG	ءْتنين	ʔətnIn	1
EA02	1	ENG	ءْتنين	ʔətnIn	1
GA01	1	ENG	ثنين	θnIn	1
GA02	1	ENG	ءْثنين	ʔəθnIn	1
LA01	1	ENG	ثنين	θnIn	1
LA02	1	ENG	تنين	tnIn	1
MA01	1	ENG	جوج	ʒUʒ	2
MA02	1	ENG	جوج	ʒUʒ	2
SA01	1	ENG	ءِثنان	ʔiθnAn	1

ID: SWADESH_024
English word: three
Context: This is number ___

EA01	1	ENG	تَلاتة	talAta	1
EA02	1	ENG	تَلاتة	talAta	1
GA01	1	ENG	ثَلاثة	θalAθa	1
GA02	1	ENG	ثَلاثة	θalAθa	1
LA01	1	ENG	ثَلاث	θalAθ	1
LA02	1	ENG	تلاتة	tlAtə	1
MA01	1	ENG	تلاتة	tlAta	1
MA02	1	ENG	تلاتة	tlAta	1
SA01	1	ENG	ثَلاثة	θalAθa	1

ID: SWADESH_025
English word: four
Context: This is number ___

EA01	1	ENG	عَربَعة	ʔarbaʕa	1
EA02	1	ENG	عَربَعة	ʔarbaʕa	1
GA01	1	ENG	عَربعة	ʔarbʕa	1
GA02	1	ENG	عَربَعة	ʔarbaʕa	1
LA01	1	ENG	عَربعة	ʔarbaʕa	1
LA02	1	ENG	عَربَعَ	ʔarbaʕa	1
LA02	2	VAR	عَربَ	ʔarba	1
MA01	1	ENG	رْبعا	rəbʕa	1
MA02	1	ENG	رْبعا	rəbʕa	1
SA01	1	ENG	عَربَعة	ʔarbaʕa	1

ID: SWADESH_026
English word: five
Context: This is number ___

EA01	1	ENG	خَمسة	xamsa	1
EA02	1	ENG	خَمسة	xamsa	1
GA01	1	ENG	خَمسة	xamsa	1
GA02	1	ENG	خَمسة	xamsa	1
LA01	1	ENG	خَمسة	xamsə	1
LA02	1	ENG	خَمسة	xamsə	1
MA01	1	ENG	خْمسة	xəmsa	1
MA02	1	ENG	خْمسة	xəmsa	1
SA01	1	ENG	خَمسة	xamsa	1

ID: SWADESH_027
English word: big
Context: a ___ book

EA01	1	ENG	كبير	kbIr	1
EA02	1	ENG	كْبير	kəbIr	1
GA01	1	ENG	كْبير	kəbIr	1

GA02	1	ENG	كَبير	kabIr	1
LA01	1	ENG	كبير	kbIr	1
LA02	1	ENG	كبير	kbIr	1
MA01	1	ENG	كُبير	kəbIr	1
MA02	1	ENG	كبير	kbIr	1
SA01	1	ENG	كَبير	kabIr	1

ID: SWADESH_028
English word: long
Context: a ___ street

EA01	1	ENG	طَويل	tˤawIl	1
EA02	1	ENG	طَويل	tˤawIl	1
GA01	1	ENG	طُويل	tˤuwIl	1
GA02	1	ENG	طَويل	tˤawIl	1
LA01	1	ENG	طويل	tˤwIl	1
LA01	2	ENG	طَويل	tˤawIl	1
LA02	1	ENG	طَويل	tˤawIl	1
MA01	1	ENG	طْويل	tˤəwIl	1
MA02	1	ENG	طويل	tˤwIl	1
SA01	1	ENG	طَويل	tˤawyl	1

ID: SWADESH_029
English word: wide
Context: a ___ street

EA01	1	ENG	واسِع	wAsiʕ	1
EA02	1	ENG	عَريض	ʕarIdˤ	2
EA02	2	ENG	واسِع	wAsiʕ	1
GA01	1	ENG	وُسيع	wusIʕ	1
GA02	1	ENG	عَريظ	ʕarIðˤ	2
LA01	1	ENG	واسِع	wAsiʕ	1
LA01	2	VAR	عَريظ	ʕarIðˤ	2
LA02	1	ENG	عَريض	ʕarIdˤ	2
LA02	2	VAR	واسِع	wAsiʕ	1
MA01	1	ENG	واسِع	wAsiʕ	1
MA02	1	ENG	واسِع	wAsiʕ	1
SA01	1	ENG	واسِع	wAsiʕ	1
SA01	2	ENG	عَريض	ʕarIdˤ	2

ID: SWADESH_030
English word: thick
Context: a ___ wooden board

EA01	1	ENG	تْخين	təxIn	1
EA02	1	ENG	تْخين	təxIn	1

EA02	2	VAR	سَميك	samɪk	2
GA01	1	ENG	سْميك	səmɪk	2
GA02	1	ENG	مَتين	matɪn	3
LA01	1	ENG	ثْخين	θxɪn	1
LA01	2	VAR	خميل	xmɪl	4
LA01	3	VAR	غَليظ	ʁalɪðˤ	5
LA02	1	ENG	سميك	smɪk	2
LA02	2	VAR	طخين	tˤxɪn	1
MA01	1	ENG	غليض	ʁlɪdˤ	5
MA01	2	VAR	سميك	smɪk	2
MA02	1	ENG	غليض	ʁlɪdˤ	5
SA01	1	ENG	سَميك	samɪk	2
SA01	2	ENG	ثَخين	θaxɪn	1
SA01	3	ENG	غَليظ	ʁalɪðˤ	5

ID: SWADESH_031
English word: heavy
Context: a ___ wooden board

EA01	1	ENG	تْءيل	təʔɪl	1
EA02	1	ENG	تْءيل	təʔɪl	1
GA01	1	ENG	ثُگيل	θəgɪl	1
GA02	1	ENG	ثَگيل	θagɪl	1
LA01	1	ENG	ثگيل	θgɪl	1
LA02	1	ENG	طءيل	tˤʔɪl	1
MA01	1	ENG	تقيل	tqɪl	1
MA02	1	ENG	تقيل	tqɪl	1
SA01	1	ENG	ثَقيل	θaqɪl	1

ID: SWADESH_032
English word: small
Context: a ___ wooden board

EA01	1	ENG	صُغَيَّر	sˤuʁay+ar	1
EA02	1	ENG	صُغَيَّر	sˤuʁay+ar	1
GA01	1	ENG	صغير	sˤʁɪr	1
GA02	1	ENG	صَغير	sˤaʁɪr	1
LA01	1	ENG	ز،غير	zˤʁɪr	1
LA02	1	ENG	ز،غير	zˤʁɪr	1
MA01	1	ENG	صغير	sˤʁɪr	1
MA02	1	ENG	صغير	sˤʁɪr	1
SA01	1	ENG	صَغير	sˤaʁɪr	1

ID: SWADESH_033
English word: short

Context: a ___ man

EA01	1	ENG	ءْصَيَّر	ʔəsˤay+ar	1
EA02	1	ENG	ءُصَيَّر	ʔusˤay+ar	1
GA01	1	ENG	گصير	gsˤIr	1
GA02	1	ENG	گصير	gasˤIr	1
LA01	1	ENG	كصير	ksˤIr	1
LA01	2	ENG	گصير	gasˤIr	1
LA02	1	ENG	ءَصير	ʔasˤIr	1
MA01	1	ENG	قصير	qsˤIr	1
MA02	1	ENG	قصير	qsˤIr	1
SA01	1	ENG	قَصير	qasˤIr	1

ID: SWADESH_034
English word: narrow
Context: a ___ street

EA01	1	ENG	دَيَّء	day+aʔ	1
EA02	1	ENG	دَيَّء	day+aʔ	1
GA01	1	ENG	ظَيِّگ	ðˤay+ig	1
GA02	1	ENG	ظَيِّگ	ðˤay+ig	1
LA01	1	ENG	ظَيِّگ	ðˤay+ig	1
LA02	1	ENG	ضَيِّء	dˤay+iʔ	1
MA01	1	ENG	ضِيِّق	dˤiy+iq	1
MA02	1	ENG	ضِيِّق	dˤiy+iq	1
SA01	1	ENG	ضَيِّق	dˤay+iq	1

ID: SWADESH_035
English word: thin
Context: a ___ wooden board

EA01	1	ENG	رُفَيَّع	rufay+aʕ	1
EA02	1	ENG	رُفَيَّع	rufay+aʕ	1
GA01	1	ENG	سَخيف	saxIf	2
GA01	2	VAR	نْحيف	nəħIf	3
GA02	1	ENG	ظَعيف	ðˤaʕIf	4
LA01	1	ENG	نحيف	nħIf	3
LA01	2	VAR	رَگيگ	raglg	5
LA02	1	ENG	نَحيف	naħIf	3
LA02	2	VAR	رفيع	rfIʕ	1
LA02	3	VAR	رءيء	rʔIʔ	5
MA01	1	ENG	رقيق	rqIq	5
MA02	1	ENG	رقيق	rqIq	5
SA01	1	ENG	رَفيع	rafIʕ	1
SA01	2	ENG	رَقيق	raqIq	5
SA01	3	ENG	نَحيف	naħIf	3

SA01	4	ENG	ضَعيف	dˤaʕIf	4

ID: SWADESH_036
English word: woman
Context: This is a ___

EA01	1	ENG	سِت	sit	1
EA02	1	ENG	سِت	sit	1
GA01	1	ENG	حُرمة	ħurma	2
GA01	2	VAR	مَرَ	mara	3
GA02	1	ENG	مَرا	mara	3
LA01	1	ENG	مَرَ	mara	3
LA02	1	ENG	مَرَ	mara	3
LA02	2	VAR	سِت	sit	1
MA01	1	ENG	مرا	mra	3
MA02	1	ENG	مرا	mra	3
SA01	1	ENG	ءِمرَءَه	ʔimraʔah	3

ID: SWADESH_037
English word: man (male)
Context: This is a ___

EA01	1	ENG	راجِل	rAgil	1
EA02	1	ENG	راجِل	rAgil	1
GA01	1	ENG	رَجّال	radʒ+Al	1
GA02	1	ENG	رَجّال	radʒ+Al	1
LA01	1	ENG	زَلَمة	zalamə	2
LA02	1	ENG	رِجّال	riʒ+Al	1
LA02	2	ENG	زَلَمة	zalamə	2
MA01	1	ENG	راجِل	rAʒil	1
MA02	1	ENG	راجِل	rAʒil	1
SA01	1	ENG	رَجُل	radʒul	1

ID: SWADESH_038
English word: man (human)
Context: This is a ___ (as opposed to other species)

EA01	1	ENG	بَنِءادَم	baniʔAdam	1
EA02	1	ENG	بَنِءادَم	baniʔAdam	1
GA01	1	ENG	ءادمي	ʔAdmi	2
GA02	1	ENG	ءِنسان	ʔinsAn	3
LA01	1	ENG	ءادمي	ʔAdmi	2
LA01	2	ENG	بَنِءادَم	baniʔAdam	1
LA02	1	ENG	ءِنسان	ʔinsAn	3
MA01	1	ENG	بنادَم	bnAdam	1
MA01	2	ENG	ءِنسان	ʔinsAn	3

MA02	1	ENG	بنادَم	bnAdam	1
SA01	1	ENG	ءِنسان	ʔinsAn	3
SA01	2	ENG	ءادَمي	ʔAdamiy	2
SA01	3	ENG	ءِبنءادَم	ʔibnʔAdam	1

ID: SWADESH_039
English word: child
Context: This is a ___ (5 years old)

EA01	1	ENG	وَلَد	walad	1
EA01	2	VAR	عَيِّل	ʕay+il	2
EA02	1	ENG	طِفل	tˁifl	3
EA02	2	VAR	عَيِّل	ʕay+il	2
GA01	1	ENG	طْفْل	tˁəfəl	3
GA01	2	VAR	جاهِل	ʤAhil	4
GA02	1	ENG	جاهِل	ʤAhil	4
LA01	1	ENG	طْفْل	tˁəfəl	3
LA02	1	ENG	وَلَد	walad	1
LA02	2	VAR	طْفْل	tˁəfəl	3
MA01	1	ENG	دِرّي	dir+i	5
MA01	2	VAR	وِلد	wild	1
MA02	1	ENG	وِلد	wild	1
MA02	2	VAR	دِرّي	dir+i	5
SA01	1	ENG	طِفل	tˁifl	3

ID: SWADESH_040
English word: wife
Context: (as pronounced in the context provided) She is the ___ of Ali

EA01	1	ENG	مِرات	mirAt	1
EA01	2	VAR	زوجة	zUga	2
EA02	1	ENG	مِرات	mirAt	1
GA01	1	ENG	زوجَت	zUʤat	2
GA02	1	ENG	زوجَت	zUʤat	2
LA01	1	ENG	مَرَت	marat	1
LA02	1	ENG	مَرت	mart	1
LA02	2	VAR	زَوجِت	zawʒit	2
MA01	1	ENG	مرات	mrAt	1
MA02	1	ENG	مرات	mrAt	1
SA01	1	ENG	زَوجَت	zawʤat	2
SA01	2	ENG	ءِمرَءَت	ʔimraʔat	1

ID: SWADESH_041
English word: husband
Context: (as pronounced in the context provided) He is the ___ of Salma

EA01	1	ENG	زوج	zUg	1
EA01	2	VAR	جوز	gUz	1
EA02	1	ENG	جوز	gUz	1
GA01	1	ENG	زوج	zUʤ	1
GA01	2	VAR	رَجِل	raʤil	2
GA02	1	ENG	زوج	zUʤ	1
GA02	2	VAR	رَجِل	raʤil	2
LA01	1	ENG	جوز	ʤUz	1
LA02	1	ENG	زوج	zUʒ	1
MA01	1	ENG	راجِل	rAʒil	2
MA02	1	ENG	راجِل	rAʒil	2
SA01	1	ENG	زَوج	zawʤ	1

ID: SWADESH_042
English word: mother
Context: (as pronounced in the context provided) She is the ___ of Ali

EA01	1	ENG	والدِت	wAldit	1
EA01	2	VAR	عُم	ʔum	2
EA02	1	ENG	عُم	ʔum	2
GA01	1	ENG	عُم	ʔum	2
GA01	2	VAR	والدَت	wAldat	1
GA02	1	ENG	عُم	ʔum	2
LA01	1	ENG	عُم	ʔum	2
LA02	1	ENG	عِم	ʔim	2
LA02	2	VAR	والدِت	wAldit	1
LA02	3	VAR	والدِه	wAldih	1
MA01	1	ENG	لوالِدا	lwAlida	1
MA01	2	VAR	مّاين	m+Ayn	2
MA01	3	VAR	ماما	mAma	2
MA02	1	ENG	مّو	m+u	2
MA02	2	ENG	ماما	mAma	2
SA01	1	ENG	والِدة	wAlida	1
SA01	2	ENG	عُم	ʔum	2

ID: SWADESH_043
English word: father
Context: (as pronounced in the context provided) He is the ___ of Ali

EA01	1	ENG	والِد	wAlid	1
EA01	2	VAR	عَبو	ʔabu	2
EA02	1	ENG	عَبو	ʔabu	2
GA01	1	ENG	عَبو	ʔabu	2
GA01	2	VAR	والِد	wAlid	1
GA02	1	ENG	عَبو	ʔabu	2

LA01	1	ENG	عَبو	ʔabu	2
LA02	1	ENG	عَبو	ʔabu	2
LA02	2	ENG	عَب	ʔab	2
LA02	3	VAR	والِد	wAlid	1
MA01	1	ENG	لوالِد	lwAlid	1
MA01	2	VAR	بّاين	b+Ayn	2
MA01	3	VAR	بابا	bAba	2
MA02	1	ENG	بّا	b+a	2
MA02	2	ENG	بابا	bAba	2
SA01	1	ENG	عَب	ʔab	2
SA01	2	ENG	والِد	wAlid	1

ID: SWADESH_044
English word: animal
Context: The elephant is a big ___

EA01	1	ENG	حَيَوان	ħayawAn	1
EA02	1	ENG	حَيَوان	ħayawAn	1
GA01	1	ENG	حَيوان	ħaywAn	1
GA02	1	ENG	حَيَوان	ħayawAn	1
LA01	1	ENG	حَيوان	ħaywAn	1
LA02	1	ENG	حَيَوان	ħayawAn	1
MA01	1	ENG	حَيَوان	ħayawAn	1
MA02	1	ENG	حَيَوان	ħayawAn	1
SA01	1	ENG	حَيَوان	ħayawAn	1

ID: SWADESH_045
English word: fish
Context: This ___ is about 20" in length, I am not sure what type it is.

EA01	1	ENG	سَمَكة	samaka	1
EA02	1	ENG	سَمَكة	samaka	1
GA01	1	ENG	سمَكة	smaka	1
GA02	1	ENG	سَمَچة	sama¢a	1
LA01	1	ENG	سَمَكة	samaka	1
LA02	1	ENG	سَمكة	samkə	1
MA01	1	ENG	حوتة	ħUta	2
MA02	1	ENG	حوتة	ħUta	2
SA01	1	ENG	سَمَكة	samaka	1

ID: SWADESH_046
English word: bird
Context: This ___ is about 5" in length, I am not sure what type it is.

EA01	1	ENG	عَصفور	ʕasˁfUr	1
EA02	1	ENG	عَصفور	ʕasˁfUr	1

GA01	1	ENG	عُصفور	ʕusʿfUr	1
GA02	1	ENG	عُصفور	ʕusʿfUr	1
LA01	1	ENG	عَصفور	ʕasʿfUr	1
LA02	1	ENG	عَصفور	ʕasʿfUr	1
MA01	1	ENG	طير	tʿIr	2
MA02	1	ENG	طير	tʿIr	2
SA01	1	ENG	عُصفور	ʕusʿfUr	1
SA01	2	ENG	طَير	tʿayr	2

ID: SWADESH_047
English word: dog
Context: This is a ___

EA01	1	ENG	كَلب	kalb	1
EA02	1	ENG	كَلب	kalb	1
GA01	1	ENG	كَلب	kalb	1
GA02	1	ENG	چَلب	ʧalb	1
LA01	1	ENG	كَلب	kalb	1
LA02	1	ENG	كَلب	kalb	1
MA01	1	ENG	كَلب	kalb	1
MA02	1	ENG	كَلب	kalb	1
SA01	1	ENG	كَلب	kalb	1

ID: SWADESH_048
English word: louse
Context: The boy has one ___ in his hair

EA01	1	ENG	ءَملة	ʔamla	1
EA02	1	ENG	ءَملة	ʔamla	1
GA01	1	ENG	گَملة	gamla	1
GA02	1	ENG	گَملة	gamla	1
LA01	1	ENG	گَملة	gamlə	1
LA02	1	ENG	ءَملة	ʔamlə	1
MA01	1	ENG	گُملة	gəmla	1
MA02	1	ENG	قَملة	qamla	1
SA01	1	ENG	قَملة	qamla	1

ID: SWADESH_049
English word: snake
Context: A poisonous ___ is in the garden (the size of a walking stick)

EA01	1	ENG	ثْعبان	təʕbAn	1
EA01	2	VAR	حَيّة	ħay+a	2
EA02	1	ENG	ثْعبان	təʕbAn	1
GA01	1	ENG	حَيّة	ħay+a	2
GA02	1	ENG	حَيّة	ħay+a	2

LA01	1	ENG	حَيّة	ħay+ə	2
LA02	1	ENG	حَيّة	ħay+ə	2
MA01	1	ENG	حْنش	ħənʃ	3
MA02	1	ENG	حْنش	ħənʃ	3
MA02	2	VAR	لفعة	lfʕa	4
SA01	1	ENG	حَيّة	ħay+a	2
SA01	2	ENG	ثُعبان	θuʕbAn	1
SA01	3	ENG	ءَفعا	ʔafʕA	4

ID: SWADESH_050
English word: worm
Context: This is a ___

EA01	1	ENG	دودة	dUda	1
EA02	1	ENG	دودة	dUda	1
GA01	1	ENG	دودة	dUda	1
GA02	1	ENG	دودة	dUda	1
LA01	1	ENG	دودة	dUdə	1
LA02	1	ENG	دودة	dUdə	1
MA01	1	ENG	دودة	dUda	1
MA02	1	ENG	دودة	dUda	1
SA01	1	ENG	دودة	dUda	1

ID: SWADESH_051
English word: tree
Context: This is a ___

EA01	1	ENG	شَجَرَ	ʃagara	1
EA02	1	ENG	شَجَرة	ʃagara	1
GA01	1	ENG	شجَرة	ʃʤara	1
GA02	1	ENG	شَجَرة	ʃadʒara	1
LA01	1	ENG	شَجَرة	ʃadʒara	1
LA02	1	ENG	شَجرَ	ʃaʒra	1
MA01	1	ENG	شَجرة	ʃaʒra	1
MA02	1	ENG	شَجرة	ʃaʒra	1
SA01	1	ENG	شَجَرة	ʃadʒara	1

ID: SWADESH_052
English word: forest
Context: This is a ___

EA01	1	ENG	غابة	ʁAba	1
EA02	1	ENG	غابة	ʁAba	1
GA01	1	ENG	غابة	ʁAba	1
GA02	1	ENG	غابة	ʁAba	1
LA01	1	ENG	غابة	ʁAba	1

LA02	1	ENG	غابة	ʁAbə	1
MA01	1	ENG	غابة	ʁAba	1
MA02	1	ENG	غابة	ʁAba	1
SA01	1	ENG	غابة	ʁAba	1

ID: SWADESH_053
English word: stick
Context: The boy is playing with a ___ (the size of a walking stick)

EA01	1	ENG	عَصاية	ʕasʿAya	1
EA02	1	ENG	عَصاية	ʕasʿAya	1
GA01	1	ENG	عَصاه	ʕasʿAh	1
GA02	1	ENG	عَصاية	ʕasʿAya	1
LA01	1	ENG	عَصاية	ʕasʿAyə	1
LA02	1	ENG	عَصاية	ʕasʿAyə	1
MA01	1	ENG	عْصا	ʕəsʿa	1
MA02	1	ENG	عْصا	ʕəsʿa	1
MA02	2	ENG	عود	ʕUd	2
SA01	1	ENG	عَصا	ʕasʿA	1
SA01	2	ENG	عود	ʕUd	2

ID: SWADESH_054
English word: fruit
Context: This is a/an ___ (apple, orange, grape, strawberry, and banana)

EA01	1	ENG	فاكهة	fAkha	1
EA02	1	ENG	فاكهة	fAkha	1
GA01	1	ENG	فاكهة	fAkha	1
GA02	1	ENG	فَواكِه	fawAkih	1
LA01	1	ENG	فَواكِه	fawAkih	1
LA02	1	ENG	فَواكِه	fawAkih	1
MA01	1	ENG	فاكِهة	fAkiha	1
MA01	2	VAR	ديسير	dIsIr	2
MA02	1	ENG	ديسِر	dIsir	2
SA01	1	ENG	فاكِهة	fAkiha	1

ID: SWADESH_055
English word: seed
Context: I saw a watermelon that had only one ___

EA01	1	ENG	بِزرة	bizra	1
EA02	1	ENG	بِزرة	bizra	1
GA01	1	ENG	طُعامه	tʿəʕAmh	2
GA02	1	ENG	حَبّة	ħab+a	3
LA01	1	ENG	بِزرة	bizrə	1
LA02	1	ENG	بِزرة	bizrə	1

MA01	1	ENG	زرّيعة	zr+Iʕa	4
MA02	1	ENG	زرّيعة	zr+Iʕa	4
SA01	1	ENG	بِذرة	biðra	1
SA01	2	ENG	حَبّة	ħab+a	3

ID: SWADESH_056
English word: leaf
Context: Only one ___ remains on the tree

EA01	1	ENG	وَرَءة	waraʔa	1
EA02	1	ENG	وَرَءة	waraʔa	1
GA01	1	ENG	وَرَقة	waraga	1
GA02	1	ENG	وَرَقة	waraqa	1
LA01	1	ENG	وَرَقة	waraga	1
LA02	1	ENG	وَرءا	warʔa	1
MA01	1	ENG	وْرقة	wərqa	1
MA02	1	ENG	وْرقة	wərqa	1
SA01	1	ENG	وَرَقة	waraqa	1

ID: SWADESH_057
English word: root
Context: The tree ___ (is large) or (goes deep in the soil)

EA01	1	ENG	جِزر	gizr	1
EA02	1	ENG	جِزر	gizr	1
GA01	1	ENG	جَذِر	ʤaðir	1
GA02	1	ENG	جَدِر	ʤadir	1
LA01	1	ENG	جْدِر	ʤəðir	1
LA01	2	VAR	شْرش	ʃərʃ	2
LA02	1	ENG	جِزِر	ʒizir	1
LA02	2	VAR	شِلِش	ʃiliʃ	2
MA01	1	ENG	جدِر	ʒdir	1
MA02	1	ENG	جدِر	ʒdir	1
SA01	1	ENG	شِرش	ʃirʃ	2
SA01	2	ENG	جَذر	ʤaðr	1

ID: SWADESH_058
English word: bark
Context: The tree is surrounded by a layer of ___ to protect it

EA01	1	ENG	لُحاء	luħAʔ	1
EA02	1	ENG	عِشرة	ʔiʃra	2
GA01	1	ENG	گلافة	glAfa	3
GA02	1	ENG	گِشرة	giʃra	2
LA01	1	ENG	لُحا	ləħa	1
LA02	1	ENG	عِشرة	ʔiʃrə	2

MA01	1	ENG	لحاف	lħAf	4
MA02	1	ENG	قشرة	qʃra	2
MA02	2	ENG	غشا	ʁʃa	5
SA01	1	ENG	لِحاء	liħAʔ	1

ID: SWADESH_059
English word: flower
Context: I don't know what kind of ___ this is

EA01	1	ENG	وَردة	warda	1
EA01	2	VAR	زَهرة	zahra	2
EA02	1	ENG	وَردة	warda	1
GA01	1	ENG	وَردة	warda	1
GA02	1	ENG	زَهرة	zahra	2
GA02	2	ENG	وَردة	warda	1
LA01	1	ENG	وَردة	wardə	1
LA02	1	ENG	زهرا	zhra	2
MA01	1	ENG	وْردة	wərda	1
MA02	1	ENG	وْردة	wərda	1
SA01	1	ENG	زَهرة	zahra	2

ID: SWADESH_060
English word: grass
Context: Wild ___ grows in forests

EA01	1	ENG	حَشيش	ħaʃlʃ	1
EA02	1	ENG	حَشيش	ħaʃlʃ	1
GA01	1	ENG	عْشب	ʕəʃb	2
GA02	1	ENG	حَشيش	ħaʃlʃ	1
LA01	1	ENG	عْشْب	ʕəʃəb	2
LA02	1	ENG	عِشِب	ʕiʃib	2
LA02	2	VAR	حَشيش	ħaʃlʃ	1
MA01	1	ENG	ربيع	rbIʕ	3
MA02	1	ENG	ربيع	rbIʕ	3
SA01	1	ENG	عُشب	ʕuʃb	2
SA01	2	ENG	حَشيش	ħaʃlʃ	1

ID: SWADESH_061
English word: rope
Context: The tree climber uses a ___

EA01	1	ENG	حَبل	ħabl	1
EA02	1	ENG	حَبل	ħabl	1
GA01	1	ENG	حَبْل	ħabəl	1
GA02	1	ENG	حَبْل	ħabəl	1
LA01	1	ENG	حَبْل	ħabəl	1

LA02	1	VAR	حَبْل	ħabəl	1
MA01	1	ENG	كوردا	kUrda	2
MA01	2	VAR	حبْل	ħbəl	1
MA01	3	VAR	قْنُبا	qən+əba	3
MA02	1	ENG	قْنُبا	qən+əba	3
MA02	2	VAR	حبْل	ħbəl	1
MA02	3	VAR	كوردا	kUrda	2
SA01	1	ENG	حَبل	ħabl	1

ID: SWADESH_062
English word: skin
Context: Africans mostly have darker
___

EA01	1	ENG	جِلد	gild	1
EA01	2	VAR	بَشرا	baʃra	2
EA02	1	ENG	جِلد	gild	1
EA02	2	VAR	بَشرا	baʃra	2
GA01	1	ENG	جِلد	ʤild	1
GA01	2	VAR	بَشرا	baʃra	2
GA02	1	ENG	جِلد	ʤild	1
LA01	1	ENG	بَشرا	baʃra	2
LA02	1	ENG	بَشرا	baʃra	2
LA02	2	VAR	جِلِد	ʒilid	1
MA01	1	ENG	جْلد	ʒəld	1
MA01	2	VAR	بشرا	bʃra	2
MA02	1	ENG	جلدا	ʒlda	1
MA02	2	ENG	بشرا	bʃra	2
SA01	1	ENG	جِلد	ʤild	1
SA01	2	ENG	بَشَرة	baʃara	2

ID: SWADESH_063
English word: meat
Context: ___ from beef is red

EA01	1	ENG	لَحم	laħm	1
EA02	1	ENG	لَحم	laħm	1
GA01	1	ENG	لَحْم	laħəm	1
GA02	1	ENG	لَحْم	laħəm	1
LA01	1	ENG	لَحْم	laħəm	1
LA02	1	ENG	لَحْم	laħəm	1
MA01	1	ENG	لحْم	lħəm	1
MA02	1	ENG	لحْم	lħəm	1
SA01	1	ENG	لَحم	laħm	1

ID: SWADESH_064
English word: blood
Context: The ___ is red

EA01	1	ENG	دَم	dam	1
EA02	1	ENG	دَم	dam	1
GA01	1	ENG	دَم	dam	1
GA02	1	ENG	دَم	dam	1
LA01	1	ENG	دَم	dam	1
LA02	1	ENG	دَم	dam	1
MA01	1	ENG	دْم	dəm	1
MA02	1	ENG	دْم	dəm	1
SA01	1	ENG	دَم	dam	1

ID: SWADESH_065
English word: bone
Context: The ___ is white

EA01	1	ENG	عَضم	ʕadˤm	1
EA02	1	ENG	عَضم	ʕadˤm	1
GA01	1	ENG	عَظْم	ʕaðˤəm	1
GA02	1	ENG	عَظْم	ʕaðˤəm	1
LA01	1	ENG	عَظْم	ʕaðˤəm	1
LA02	1	ENG	عَضْم	ʕadˤəm	1
MA01	1	ENG	عضْم	ʕdˤəm	1
MA02	1	ENG	عظْم	ʕðˤəm	1
SA01	1	ENG	عَظم	ʕaðˤm	1

ID: SWADESH_066
English word:
fat
Context: The ___ is white

EA01	1	ENG	سْمين	səmIn	1
EA02	1	ENG	دِهن	dihn	2
EA02	2	VAR	سْمين	səmIn	1
GA01	1	ENG	شَحمة	ʃaħma	3
GA01	2	ENG	شَحْم	ʃaħəm	3
GA02	1	ENG	شَحْم	ʃaħəm	3
LA01	1	ENG	دْهْن	dəhən	2
LA02	1	ENG	دْهْن	dəhən	2
MA01	1	ENG	شحمة	ʃħma	3
MA02	1	ENG	شحمة	ʃħma	3
SA01	1	ENG	دُهن	duhn	2
SA01	2	ENG	شَحم	ʃaħm	3

ID: SWADESH_067
English word: egg
Context: The chicken laid an ___

EA01	1	ENG	بيضة	bIdˤa	1
EA02	1	ENG	بيضة	bIdˤa	1
GA01	1	ENG	بيظة	bIðˤa	1
GA02	1	ENG	بيظة	bIðˤa	1
LA01	1	ENG	بيظة	bIðˤa	1
LA02	1	ENG	بَيضا	baydˤa	1
MA01	1	ENG	بيضة	bIdˤa	1
MA02	1	ENG	بيضة	bIdˤa	1
SA01	1	ENG	بَيضة	baydˤa	1

ID: SWADESH_068
English word: horn
Context: The bull has a broken ___

EA01	1	ENG	ءَرن	ʔarn	1
EA02	1	ENG	ءَرن	ʔarn	1
GA01	1	ENG	گرن	garn	1
GA02	1	ENG	گرن	garn	1
LA01	1	ENG	گُرْن	gərən	1
LA02	1	ENG	ءُرْن	ʔərən	1
MA01	1	ENG	گُرن	gərn	1
MA02	1	ENG	گُرن	gərn	1
SA01	1	ENG	قَرن	qarn	1

ID: SWADESH_069
English word: tail
Context: The fox has a nice ___

EA01	1	ENG	ديل	dIl	1
EA02	1	ENG	ديل	dIl	1
GA01	1	ENG	ذيل	ðIl	1
GA02	1	ENG	دَنَب	danab	2
GA02	2	VAR	ديل	dIl	1
LA01	1	ENG	ذيل	ðIl	1
LA02	1	ENG	دَنَب	danab	2
MA01	1	ENG	زْنطيط	zəntˤItˤ	3
MA01	2	VAR	ديل	dIl	1
MA02	1	ENG	ديل	dIl	1
MA02	2	ENG	شوال	ʃwAl	4
MA02	3	VAR	زْنطيط	zəntˤItˤ	3
SA01	1	ENG	ذَيل	ðayl	1
SA01	2	ENG	ذَنَب	ðanab	2

ID: SWADESH_070
English word: feather
Context: The bird dropped a nice ___

EA01	1	ENG	ريشة	rIʃa	1
EA02	1	ENG	ريشة	rIʃa	1
GA01	1	ENG	ريشة	rIʃa	1
GA02	1	ENG	ريشة	rIʃa	1
LA01	1	ENG	ريشة	rIʃə	1
LA02	1	ENG	ريشة	rIʃə	1
MA01	1	ENG	ريشة	rIʃa	1
MA02	1	ENG	ريشة	rIʃa	1
SA01	1	ENG	ريشة	rIʃa	1

ID: SWADESH_071
English word: hair
Context: Human ___ comes in different colors (singlular)

EA01	1	ENG	شَعرة	ʃaʕra	1
EA02	1	ENG	شَعرة	ʃaʕra	1
GA01	1	ENG	شَعرة	ʃaʕra	1
GA02	1	ENG	شَعرة	ʃaʕra	1
LA01	1	ENG	شَعرا	ʃaʕra	1
LA02	1	ENG	شَعرا	ʃaʕra	1
MA01	1	ENG	شْعرة	ʃəʕra	1
MA02	1	ENG	زُغبة	zəʁba	2
MA02	2	ENG	شْعرة	ʃəʕra	1
SA01	1	ENG	شَعرة	ʃaʕra	1

ID: SWADESH_072
English word: head
Context: The cheetah's ___ is small

EA01	1	ENG	راس	rAs	1
EA01	2	ENG	دِماغ	dimAʁ	2
EA02	1	ENG	راس	rAs	1
EA02	2	VAR	دِماغ	dimAʁ	2
GA01	1	ENG	راس	rAs	1
GA02	1	ENG	راس	rAs	1
LA01	1	ENG	راس	rAs	1
LA02	1	ENG	راس	rAs	1
MA01	1	ENG	راس	rAs	1
MA02	1	ENG	راس	rAs	1
SA01	1	ENG	رَءس	raʔs	1

ID: SWADESH_073
English word: ear
Context: The boy has an ___

EA01	1	ENG	وِدن	widn	1
EA02	1	ENG	وِدن	widn	1
GA01	1	ENG	ءُذُن	ʔəðən	1
GA02	1	ENG	ءَدون	ʔadUn	1
LA01	1	ENG	ءِذُن	ʔiðən	1
LA02	1	ENG	دَينة	daynə	1
MA01	1	ENG	وْدِن	wədin	1
MA02	1	ENG	وْدِن	wədin	1
SA01	1	ENG	ءُذُن	ʔuðun	1

ID: SWADESH_074
English word: eye
Context: The boy has an ___

EA01	1	ENG	عين	ʕIn	1
EA02	1	ENG	عين	ʕIn	1
GA01	1	ENG	عين	ʕIn	1
GA02	1	ENG	عيون	ʕyUn	1
LA01	1	ENG	عين	ʕIn	1
LA02	1	ENG	عين	ʕIn	1
MA01	1	ENG	عين	ʕIn	1
MA02	1	ENG	عين	ʕIn	1
SA01	1	ENG	عَين	ʕayn	1

ID: SWADESH_075
English word: nose
Context: The boy has a
___

EA01	1	ENG	مَناخير	manAxIr	1
EA02	1	ENG	مَناخير	manAxIr	1
GA01	1	ENG	خَشِم	xaʃim	2
GA02	1	ENG	خَشِم	xaʃim	2
LA01	1	ENG	خُشْم	xəʃəm	2
LA02	1	ENG	مْنخار	mənxAr	1
LA02	2	ENG	ءَنف	ʔanf	3
MA01	1	ENG	نيف	nIf	3
MA01	2	VAR	مْنخار	mənxAr	1
MA02	1	ENG	نيف	nIf	3
MA02	2	VAR	مْنخار	mənxAr	1
SA01	1	ENG	ءَنف	ʔanf	3
SA01	2	ENG	مُنخَر	munxar	1

ID: SWADESH_076
English word: mouth
Context: The boy has a
___

EA01	1	ENG	بُء	buʔ	1
EA02	1	ENG	بُء	buʔ	1
GA01	1	ENG	فَم	fam	2
GA02	1	ENG	بوز	bUz	3
LA01	1	ENG	ثُم	θəm	4
LA02	1	ENG	ثُم	təm	4
MA01	1	ENG	فُم	fum	2
MA02	1	ENG	فُم	fum	2
SA01	1	ENG	فَم	fam	2

ID: SWADESH_077
English word: tooth
Context: The boy has a ___ (referring to the incisors)

EA01	1	ENG	سِنّة	sin+a	1
EA02	1	ENG	سِنّة	sin+a	1
GA01	1	ENG	سِن	sin	1
GA01	2	VAR	ظِرس	ðˤirs	2
GA02	1	ENG	ضَرس	dˤars	2
LA01	1	ENG	سِن	sin	1
LA02	1	ENG	سِن	sin	1
MA01	1	ENG	سنّة	sn+a	1
MA02	1	ENG	سنّة	sn+a	1
SA01	1	ENG	سِنّ	sin+	1
SA01	2	ENG	ضِرس	dˤirs	2

ID: SWADESH_078
English word: tongue
Context: The boy has a
___

EA01	1	ENG	لِسان	lisAn	1
EA02	1	ENG	لِسان	lisAn	1
GA01	1	ENG	لسان	lsAn	1
GA02	1	ENG	لسان	lsAn	1
LA01	1	ENG	لسان	lsAn	1
LA02	1	ENG	لسان	lsAn	1
MA01	1	ENG	لسان	lsAn	1
MA02	1	ENG	لسان	lsAn	1
SA01	1	ENG	لِسان	lisAn	1

ID: SWADESH_079
English word: fingernail
Context: The boy has a
___

EA01	1	ENG	ضِفر	dˤifr	1
EA02	1	ENG	ضْفر	dˤəfr	1
GA01	1	ENG	ظُفْر	ðˤufər	1
GA02	1	ENG	ظُفْر	ðˤufər	1
LA01	1	ENG	ءِظفَر	ʔiðˤfar	1
LA02	1	ENG	ضْفْر	dˤəfər	1
MA01	1	ENG	دفْر	dfər	1
MA02	1	ENG	دفْر	dfər	1
SA01	1	ENG	ظُفر	ðˤufr	1

ID: SWADESH_080
English word: foot
Context: The boy has a
___

EA01	1	ENG	رِجل	rigl	1
EA02	1	ENG	رِجل	rigl	1
GA01	1	ENG	رِجِل	riʤil	1
GA02	1	ENG	رِجِل	riʤil	1
LA01	1	ENG	رِجِل	riʤil	1
LA01	2	VAR	ءِجِر	ʔiʒir	2
LA02	1	ENG	ءِجْر	ʔiʒər	2
MA01	1	ENG	رجْل	rʒəl	1
MA02	1	ENG	رجْل	rʒəl	1
SA01	1	ENG	قَدَم	qadam	3

ID: SWADESH_081
English word: leg
Context: The boy has a
___

EA01	1	ENG	رِجل	rigl	1
EA02	1	ENG	رِجل	rigl	1
GA01	1	ENG	رِجِل	riʤil	1
GA02	1	ENG	رِجِل	riʤil	1
LA01	1	ENG	رِجِل	riʤil	1
LA01	2	VAR	ءِجِر	ʔiʒir	2
LA02	1	ENG	ءِجْر	ʔiʒər	2
MA01	1	ENG	رجْل	rʒəl	1
MA02	1	ENG	رجْل	rʒəl	1

SA01	1	ENG	ساق	sAq	3
SA01	2	ENG	رِجل	ridʒl	1

ID: SWADESH_082
English word: knee
Context: The boy has a
___

EA01	1	ENG	رُكبة	rukba	1
EA02	1	ENG	رُكبة	rukba	1
GA01	1	ENG	رْكبة	rəkba	1
GA02	1	ENG	رِكبة	rikba	1
LA01	1	ENG	رُكبة	rukba	1
LA02	1	ENG	رِكبة	rikbə	1
MA01	1	ENG	رْكبة	rəkba	1
MA02	1	ENG	رْكبة	rəkba	1
SA01	1	ENG	رُكبة	rukba	1

ID: SWADESH_083
English word: hand
Context: The boy has a
___

EA01	1	ENG	ءيد	ʔId	1
EA02	1	ENG	ءيد	ʔId	1
GA01	1	ENG	يَد	yad	1
GA02	1	ENG	ءيد	ʔId	1
LA01	1	ENG	ءيد	ʔId	1
LA02	1	ENG	ءيد	ʔId	1
MA01	1	ENG	يْد	yəd	1
MA02	1	ENG	يْد	yəd	1
SA01	1	ENG	يَد	yad	1

ID: SWADESH_084
English word: wing
Context: The bird has a
___

EA01	1	ENG	جِناح	ginAħ	1
EA02	1	ENG	جِناح	ginAħ	1
GA01	1	ENG	جَناح	dʒanAħ	1
GA02	1	ENG	جَناح	dʒanAħ	1
LA01	1	ENG	جناح	ʒnAħ	1
LA02	1	ENG	جناح	ʒnAħ	1
MA01	1	ENG	جناح	ʒnAħ	1
MA02	1	ENG	جناح	ʒnAħ	1

SA01	1	ENG	جَناح	ʤanAħ	1

ID: SWADESH_085
English word: belly
Context: The boy has a ___ (the boy is slim)

EA01	1	ENG	بَطن	batˁn	1
EA02	1	ENG	بَطن	batˁn	1
GA01	1	ENG	بَطْن	batˁən	1
GA02	1	ENG	دَبّه	dab+ah	2
GA02	2	ENG	بَطْن	batˁən	1
LA01	1	ENG	بَطْن	batˁən	1
LA02	1	ENG	بَطْن	batˁən	1
MA01	1	ENG	كْرش	kərʃ	3
MA02	1	ENG	كْرش	kərʃ	3
SA01	1	ENG	بَطن	batˁn	1

ID: SWADESH_086
English word: guts
Context: The ___ of a cow are big (Everything in the abdomen)

EA01	1	ENG	ءَحشاء	ʔaħʃAʔ	1
EA02	1	ENG	مَصارين	masˁArIn	2
GA01	1	ENG	ءَمعاء	ʔamʕAʔ	3
GA01	2	VAR	مَصارين	masˁArIn	2
GA02	1	ENG	مَصارين	masˁArIn	2
LA01	1	ENG	مَصَرين	masˁarIn	2
LA02	1	ENG	مصارين	msˁArIn	2
MA01	1	ENG	مصارِن	msˁArin	2
MA02	1	ENG	مصارِن	msˁArin	2
SA01	1	ENG	ءَحشاء	ʔaħʃAʔ	1

ID: SWADESH_087
English word: neck
Context: The boy has a
___

EA01	1	ENG	رَءَبة	raʔaba	1
EA02	1	ENG	رَءَبة	raʔaba	1
GA01	1	ENG	رگُبة	rguba	1
GA02	1	ENG	رَگبة	ragaba	1
LA01	1	ENG	رَگبة	ragaba	1
LA02	1	ENG	رَءبة	raʔbə	1
MA01	1	ENG	عْنق	ʕənq	2
MA02	1	ENG	عَنق	ʕanq	2
SA01	1	ENG	رَقَبة	raqaba	1

SA01	2	ENG	عُنق	ʕunq	2

ID: SWADESH_088
English word: back
Context: The boy has a
___

EA01	1	ENG	ضَهر	dʿahr	1
EA02	1	ENG	ضَهر	dʿahr	1
GA01	1	ENG	ظَهَر	ðʿahar	1
GA02	1	ENG	ظَهَر	ðʿahar	1
LA01	1	ENG	ظَهْر	ðʿahər	1
LA02	1	ENG	ضَهِر	dʿahir	1
MA01	1	ENG	ضهْر	dʿhər	1
MA02	1	ENG	ضهَر	dʿhar	1
SA01	1	ENG	ظَهر	ðʿahr	1

ID: SWADESH_089
English word: breast
Context: The mom is feeding the baby from her ___

EA01	1	ENG	سِدر	sidr	1
EA02	1	ENG	سِدر	sidr	1
GA01	1	ENG	صَدِر	sʿadir	1
GA02	1	ENG	صَدِر	sʿadir	1
LA01	1	ENG	صْدْر	sʿədər	1
LA02	1	ENG	صِدِر	sʿidir	1
MA01	1	ENG	صدْر	sʿdər	1
MA02	1	ENG	صدْر	sʿdər	1
SA01	1	ENG	صَدر	sʿadr	1

ID: SWADESH_090
English word: heart
Context: The ___ beats continuously

EA01	1	ENG	ءَلب	ʔalb	1
EA02	1	ENG	ءَلب	ʔalb	1
GA01	1	ENG	گَلب	galb	1
GA02	1	ENG	قَلب	qalb	1
LA01	1	ENG	گَلب	galb	1
LA02	1	ENG	ءَلْب	ʔaləb	1
MA01	1	ENG	قَلب	qalb	1
MA02	1	ENG	قَلب	qalb	1
SA01	1	ENG	قَلب	qalb	1

ID: SWADESH_091

English word: liver
Context: The ___ filters the blood

EA01	1	ENG	كِبد	kibd	1
EA02	1	ENG	كِبد	kibd	1
GA01	1	ENG	كَبْد	kabəd	1
GA02	1	ENG	چَبْد	¢abəd	1
LA01	1	ENG	كُبْد	kəbəd	1
LA02	1	ENG	كُبْد	kəbəd	1
MA01	1	ENG	كُبدة	kəbda	1
MA02	1	ENG	كُبدة	kəbda	1
SA01	1	ENG	كَبِد	kabid	1

ID: SWADESH_092
English word: to drink
Context: Past tense form: drank. The boy ___ all the water

EA01	1	ENG	شِرِب	ʃirib	1
EA02	1	ENG	شِرِب	ʃirib	1
GA01	1	ENG	شِرِب	ʃirib	1
GA02	1	ENG	شَرَب	ʃarab	1
LA01	1	ENG	شِرْب	ʃirəb	1
LA02	1	ENG	شِرِب	ʃirib	1
MA01	1	ENG	شرْب	ʃrəb	1
MA02	1	ENG	شرْب	ʃrəb	1
SA01	1	ENG	شَرِب	ʃarib	1

ID: SWADESH_093
English word: to eat
Context: Past tense form: ate. The boy ___ all the food

EA01	1	ENG	كَل	kal	1
EA02	1	ENG	كَل	kal	1
GA01	1	ENG	كَل	kal	1
GA01	2	ENG	ءَكَل	ʔakal	1
GA02	1	ENG	ءَكَل	ʔakal	1
LA01	1	ENG	ءَكَل	ʔakal	1
LA02	1	ENG	ءَكَل	ʔakal	1
MA01	1	ENG	كلا	kla	1
MA02	1	ENG	كلا	kla	1
SA01	1	ENG	ءَكَل	ʔakal	1

ID: SWADESH_094
English word: to bite
Context: Past tense form: bit. The boy ___ his little brother

EA01	1	ENG	عَض	ʕadˤ	1
EA02	1	ENG	عَض	ʕadˤ	1
GA01	1	ENG	عَظ	ʕaðˤ	1
GA02	1	ENG	عَض	ʕadˤ	1
LA01	1	ENG	عَظ	ʕaðˤ	1
LA02	1	ENG	عَض	ʕadˤ	1
MA01	1	ENG	عْض	ʕədˤ	1
MA02	1	ENG	عْض	ʕədˤ	1
SA01	1	ENG	عَضّ	ʕadˤ+	1

ID: SWADESH_095
English word: to suck
Context: Past tense form: sucked. The baby ___ the bottle

EA01	1	ENG	مَص	masˤ	1
EA02	1	ENG	مَص	masˤ	1
GA01	1	ENG	مَص	masˤ	1
GA02	1	ENG	مَص	masˤ	1
LA01	1	ENG	مَص	masˤ	1
LA02	1	ENG	مَص	masˤ	1
MA01	1	ENG	مْص	məsˤ	1
MA02	1	ENG	مْص	məsˤ	1
SA01	1	ENG	مَصّ	masˤ+	1

ID: SWADESH_096
English word: to spit
Context: Past tense form: spat. The boy ___ on the floor

EA01	1	ENG	تَف	taf	1
EA02	1	ENG	تَف	taf	1
GA01	1	ENG	تَفَل	tafal	1
GA02	1	ENG	تَفَل	tafal	1
LA01	1	ENG	تَف	taf	1
LA01	2	VAR	بَزَگ	bazag	2
LA02	1	ENG	بَزء	bazˤaʔ	2
MA01	1	ENG	دفْل	dfəl	1
MA01	2	VAR	بزَق	bzaq	2
MA02	1	ENG	بزَق	bzaq	2
MA02	2	VAR	دفْل	dfəl	1
SA01	1	ENG	بَصَق	basˤaq	2

ID: SWADESH_097
English word: to vomit
Context: Past tense form: vomited. The boy ___ on the floor

EA01	1	ENG	رَجَّع	rag+aʕ	1
EA01	2	VAR	ءِتءاية	ʔitʔAya	2
EA02	1	ENG	رَجَّع	rag+aʕ	1
EA02	2	VAR	ءِتءاية	ʔitʔAya	2
GA01	1	ENG	رَجَّع	radʒ+aʕ	1
GA01	2	VAR	ستَفرَغ	stafraʁ	3
GA02	1	ENG	زَع	zaʕ	4
LA01	1	ENG	راجَع	rAdʒaʕ	1
LA02	1	ENG	تفَوَّع	tfaw+aʕ	5
LA02	2	VAR	ستَفرَغ	stafraʁ	3
LA02	3	VAR	راجَع	rAʒaʕ	1
MA01	1	ENG	رْد	rəd	6
MA02	1	ENG	رْد	rəd	6
MA02	2	VAR	رَجَّع	raʒ+aʕ	1
SA01	1	ENG	تَقَيَّء	taqay+aʔ	2
SA01	2	ENG	ءِستَفرَغ	ʔistafraʁ	3

ID: SWADESH_098
English word: to blow
Context: Past tense form: blew. The boy ___ in the balloon

EA01	1	ENG	نَفَخ	nafax	1
EA02	1	ENG	نَفَخ	nafax	1
GA01	1	ENG	نْفَخ	nəfax	1
GA02	1	ENG	نَفَخ	nafax	1
LA01	1	ENG	نَفَخ	nafax	1
LA02	1	ENG	نَفَخ	nafax	1
MA01	1	ENG	نفْخ	nfəx	1
MA02	1	ENG	نفْخ	nfəx	1
SA01	1	ENG	نَفَخ	nafax	1

ID: SWADESH_099
English word: to breathe
Context: Past tense form: breathed. The boy ___ the air

EA01	1	ENG	ءِتنَفِّس	ʔitnaf+is	1
EA02	1	ENG	ءِتنَفِّس	ʔitnaf+is	1
GA01	1	ENG	تْنَفَّس	tənaf+as	1
GA02	1	ENG	تنَفَّس	tnaf+as	1
LA01	1	ENG	تنَفَّس	tnaf+as	1
LA02	1	ENG	تنَفَّس	tnaf+as	1
MA01	1	ENG	تنْفُّس	tnəf+əs	1
MA02	1	ENG	تنْفُّس	tnəf+əs	1
SA01	1	ENG	تَنَفَّس	tanaf+as	1

ID: SWADESH_100
English word: to laugh
Context: Past tense form: laughed. The boy ___ at the scene

EA01	1	ENG	دِحِك	diħik	1
EA02	1	ENG	دِحِك	diħik	1
GA01	1	ENG	ظِحِك	ðˤiħik	1
GA02	1	ENG	ظَحَك	ðˤaħak	1
LA01	1	ENG	ظِحِك	ðˤiħik	1
LA02	1	ENG	ضِحِك	dˤiħik	1
LA02	2	ENG	ضَّحَّك	dˤ+aħ+ak	1
MA01	1	ENG	ضحْك	dˤħək	1
MA02	1	ENG	ضحْك	dˤħək	1
SA01	1	ENG	ضَحِك	dˤaħik	1

ID: SWADESH_101
English word: to see
Context: Past tense form: saw. The boy ___ the tree

EA01	1	ENG	شاف	ʃAf	1
EA02	1	ENG	شاف	ʃAf	1
GA01	1	ENG	شاف	ʃAf	1
GA02	1	ENG	چاف	ʧAf	1
LA01	1	ENG	شاف	ʃAf	1
LA02	1	ENG	شاف	ʃAf	1
MA01	1	ENG	شاف	ʃAf	1
MA02	1	ENG	شاف	ʃAf	1
SA01	1	ENG	رَءا	raʔA	2

ID: SWADESH_102
English word: to hear
Context: Past tense form: heard. The boy ___ the music

EA01	1	ENG	سِمِع	simiʕ	1
EA02	1	ENG	سِمِع	simiʕ	1
GA01	1	ENG	سِمِع	simiʕ	1
GA02	1	ENG	سَمَع	samaʕ	1
LA01	1	ENG	سِمِع	simiʕ	1
LA02	1	ENG	سِمِع	simiʕ	1
MA01	1	ENG	سمْع	səmʕ	1
MA02	1	ENG	سمْع	səmʕ	1
SA01	1	ENG	سَمِع	samiʕ	1

ID: SWADESH_103
English word: to know
Context: Past tense form: knew. The boy ___ the toy

EA01	1	ENG	عِرِف	ʕirif	1
EA02	1	ENG	عِرِف	ʕirif	1
GA01	1	ENG	عَرَف	ʕaraf	1
GA02	1	ENG	عَرَف	ʕaraf	1
LA01	1	ENG	عِرِف	ʕirif	1
LA02	1	ENG	عِرِف	ʕirif	1
MA01	1	ENG	عرْف	ʕrəf	1
MA02	1	ENG	عرْف	ʕrəf	1
SA01	1	ENG	عَرَف	ʕaraf	1

ID: SWADESH_104
English word: to think
Context: Past tense form: thought. The boy ___ of a question

EA01	1	ENG	فَكَّر	fak+ar	1
EA02	1	ENG	فَكَّر	fak+ar	1
GA01	1	ENG	فَكَّر	fak+ar	1
GA02	1	ENG	فَكَّر	fak+ar	1
LA01	1	ENG	فَكَّر	fak+ar	1
LA02	1	ENG	فَكَّر	fak+ar	1
MA01	1	ENG	فكُّر	fk+ər	1
MA02	1	ENG	فكُّر	fk+ər	1
SA01	1	ENG	فَكَّر	fak+ar	1

ID: SWADESH_105
English word: to smell
Context: Past tense form: smelled. The boy ___ the rose

EA01	1	ENG	شَم	ʃam	1
EA02	1	ENG	شَم	ʃam	1
GA01	1	ENG	شَم	ʃam	1
GA02	1	ENG	شَم	ʃam	1
LA01	1	ENG	شَم	ʃam	1
LA02	1	ENG	شَم	ʃam	1
MA01	1	ENG	شْم	ʃəm	1
MA02	1	ENG	شْم	ʃəm	1
SA01	1	ENG	شَمّ	ʃam+	1

ID: SWADESH_106
English word: to fear
Context: Past tense form: feared. The boy ___ the dog

EA01	1	ENG	خاف	xAf	1
EA02	1	ENG	خاف	xAf	1
GA01	1	ENG	خاف	xAf	1
GA02	1	ENG	خاف	xAf	1

LA01	1	ENG	خاف	xAf	1
LA02	1	ENG	خاف	xAf	1
MA01	1	ENG	خاف	xAf	1
MA02	1	ENG	خاف	xAf	1
SA01	1	ENG	خاف	xAf	1

ID: SWADESH_107
English word: to sleep
Context: Past tense form: slept. The boy ___

EA01	1	ENG	نام	nAm	1
EA02	1	ENG	نام	nAm	1
GA01	1	ENG	نام	nAm	1
GA01	2	VAR	رِگد	rigad	2
GA02	1	ENG	نام	nAm	1
LA01	1	ENG	نام	nAm	1
LA02	1	ENG	نام	nAm	1
MA01	1	ENG	نعْس	nʕəs	3
MA02	1	ENG	نعْس	nʕəs	3
SA01	1	ENG	نام	nAm	1
SA01	2	ENG	رَقَد	raqad	2

ID: SWADESH_108
English word: to live
Context: Past tense form: lived. The hero ___ four centuries ago

EA01	1	ENG	عاش	ʕAʃ	1
EA02	1	ENG	عاش	ʕAʃ	1
GA01	1	ENG	عاش	ʕAʃ	1
GA02	1	ENG	عاش	ʕAʃ	1
LA01	1	ENG	عاش	ʕAʃ	1
LA02	1	ENG	عاش	ʕAʃ	1
MA01	1	ENG	عاش	ʕAʃ	1
MA02	1	ENG	عاش	ʕAʃ	1
SA01	1	ENG	عاش	ʕAʃ	1

ID: SWADESH_109
English word: to die
Context: Past tense form: died. The hero ___ four centuries ago

EA01	1	ENG	مات	mAt	1
EA01	2	VAR	توَفّا	twaf+a	2
EA02	1	ENG	مات	mAt	1
EA02	2	VAR	توَفّا	twaf+a	2
GA01	1	ENG	مات	mAt	1
GA01	2	VAR	توَفّا	twaf+a	2

GA02	1	ENG	مات	mAt	1
LA01	1	ENG	مات	mAt	1
LA01	2	VAR	توَفّا	twaf+a	2
LA02	1	ENG	مات	mAt	1
LA02	2	ENG	توَفّا	twaf+a	2
MA01	1	ENG	مات	mAt	1
MA01	2	VAR	توُفّا	twuf+a	2
MA02	1	ENG	مات	mAt	1
MA02	2	VAR	توُفّا	twuf+a	2
SA01	1	ENG	مات	mAt	1
SA01	2	ENG	تَوَفّا	tawaf+A	2

ID: SWADESH_110
English word: to kill
Context: Past tense form: killed. The thief ___ the boy

EA01	1	ENG	ءَتَل	ʔatal	1
EA02	1	ENG	ءَتَل	ʔatal	1
GA01	1	ENG	گِتَل	gital	1
GA01	2	ENG	ذْبَح	ðəbaħ	2
GA02	1	ENG	قَتَل	qatal	1
LA01	1	ENG	گَتَل	gatal	1
LA01	2	VAR	ذَبَح	ðabaħ	2
LA02	1	ENG	ءَتَل	ʔatal	1
MA01	1	ENG	قتْل	qtəl	1
MA02	1	ENG	قتْل	qtəl	1
SA01	1	ENG	قَتَل	qatal	1
SA01	2	ENG	ذَبَح	ðabaħ	2

ID: SWADESH_111
English word: to fight
Context: Past tense form: fought. The boy ___ with his friend

EA01	1	ENG	تخانْء	txAnəʔ	1
EA02	1	ENG	تخانْء	txAnəʔ	1
GA01	1	ENG	تْهاوَش	təhAwaʃ	2
GA01	2	VAR	تْظارَب	təðˁArab	3
GA02	1	ENG	تهاوَش	thAwaʃ	2
LA01	1	ENG	تهاوَش	thAwaʃ	2
LA02	1	ENG	ءاتَل	ʔAtal	4
LA02	2	VAR	تخانَء	txAnaʔ	1
MA01	1	ENG	دّابْز	d+Abəz	5
MA02	1	ENG	دّابْز	d+Abəz	5
MA02	2	VAR	ضّارْب	dˁ+Arəb	3
SA01	1	ENG	تَقاتَل	taqAtal	4

ID: SWADESH_112
English word: to hunt
Context: Past tense form: hunted. The man ___ in the forest

EA01	1	ENG	صطاد	sˤtˤAd	1
EA02	1	ENG	صطاد	sˤtˤAd	1
GA01	1	ENG	صاد	sˤAd	1
GA02	1	ENG	صاد	sˤAd	1
LA01	1	ENG	تصَيَّد	tsˤay+ad	1
LA02	1	ENG	صَيَّد	sˤay+ad	1
LA02	2	VAR	صاد	sˤAd	1
MA01	1	ENG	صْيُّد	sˤəy+əd	1
MA02	1	ENG	صْيُّد	sˤəy+əd	1
SA01	1	ENG	صاد	sˤAd	1
SA01	2	ENG	ءِصطاد	ʔisˤtˤAd	1

ID: SWADESH_113
English word: to hit
Context: Past tense form: hit. The man ___ the boy

EA01	1	ENG	ضَرَب	dˤarab	1
EA02	1	ENG	ضَرَب	dˤarab	1
GA01	1	ENG	ظَرَب	ðˤarab	1
GA01	2	ENG	طَگ	tˤag	2
GA02	1	ENG	ظَرَب	ðˤarab	1
LA01	1	ENG	ضَرَب	dˤarab	1
LA02	1	ENG	ضَرَب	dˤarab	1
MA01	1	ENG	ضرْب	dˤrəb	1
MA02	1	ENG	ضرْب	dˤrəb	1
SA01	1	ENG	ضَرَب	dˤarab	1

ID: SWADESH_114
English word: to cut
Context: Past tense form: cut. The boy ___ the sandwitch with the knife

EA01	1	ENG	ءَطَع	ʔatˤaʕ	1
EA02	1	ENG	ءَطَع	ʔatˤaʕ	1
GA01	1	ENG	گص	gasˤ	2
GA02	1	ENG	گص	gasˤ	2
LA01	1	ENG	گطَع	gatˤaʕ	1
LA02	1	ENG	ءَطَع	ʔatˤaʕ	1
MA01	1	ENG	قطَع	qtˤaʕ	1
MA02	1	ENG	قطَع	qtˤaʕ	1
SA01	1	ENG	قَطَع	qatˤaʕ	1
SA01	2	ENG	قَصّ	qasˤ+	2

ID: SWADESH_115
English word: to split
Context: Past tense form: split. The boy ___ the cake into two equal pieces

EA01	1	ENG	ءَسَم	ʔasam	1
EA02	1	ENG	ءَسَم	ʔasam	1
GA01	1	ENG	گِسَم	gisam	1
GA02	1	ENG	گَسَم	gasam	1
LA01	1	ENG	گَسَم	gasam	1
LA02	1	ENG	ءَصَم	ʔasˤam	1
MA01	1	ENG	فرَق	fraq	2
MA01	2	VAR	قسْم	qsəm	1
MA02	1	ENG	قسْم	qsəm	1
SA01	1	ENG	قَسَم	qasam	1

ID: SWADESH_116
English word: to stab
Context: Past tense form: stabbed. He ___ the man in the heart

EA01	1	ENG	غَز	ʁaz	1
EA02	1	ENG	طَعَن	tˤaʕan	2
EA02	2	VAR	غَز	ʁaz	1
GA01	1	ENG	طَعَن	tˤaʕan	2
GA02	1	ENG	طَعَن	tˤaʕan	2
LA01	1	ENG	طَعَن	tˤaʕan	2
LA02	1	ENG	طَعَن	tˤaʕan	2
LA02	2	ENG	غَز	ʁaz	1
MA01	1	ENG	طعْن	tˤʕən	2
MA02	1	ENG	طعْن	tˤʕən	2
SA01	1	ENG	طَعَن	tˤaʕan	2

ID: SWADESH_117
English word: to scratch
Context: Past tense form: scratched. The boy ___ his back

EA01	1	ENG	هَرَش	haraʃ	1
EA02	1	ENG	هَرَش	haraʃ	1
GA01	1	ENG	حَك	ħak	2
GA02	1	ENG	حَك	ħak	2
LA01	1	ENG	حَك	ħak	2
LA02	1	ENG	حَك	ħak	2
MA01	1	ENG	حك	ħk	2
MA02	1	ENG	حْك	ħək	2
SA01	1	ENG	حَكّ	ħak+	2

ID: SWADESH_118
English word: to dig
Context: Past tense form: dug. The boy ___ a hole at the beach

EA01	1	ENG	حَفَر	ħafar	1
EA02	1	ENG	حَفَر	ħafar	1
GA01	1	ENG	حَفَر	ħafar	1
GA02	1	ENG	حَفَر	ħafar	1
LA01	1	ENG	حَفَر	ħafar	1
LA02	1	ENG	حَفَر	ħafar	1
MA01	1	ENG	حفَر	ħfar	1
MA02	1	ENG	حفَر	ħfar	1
SA01	1	ENG	حَفَر	ħafar	1

ID: SWADESH_119
English word: to swim
Context: Past tense form: swam. The boy ___ across the river

EA01	1	ENG	عام	ʕAm	1
EA02	1	ENG	عام	ʕAm	1
GA01	1	ENG	سِبَح	sibaħ	2
GA02	1	ENG	سَبَح	sabaħ	2
LA01	1	ENG	سَبَح	sabaħ	2
LA02	1	ENG	تسَبَّح	tsab+aħ	2
LA02	2	VAR	سَبَح	sabaħ	2
LA02	3	VAR	سِبِح	sibiħ	2
MA01	1	ENG	عام	ʕAm	1
MA02	1	ENG	عام	ʕAm	1
SA01	1	ENG	سَبَح	sabaħ	2
SA01	2	ENG	عام	ʕAm	1

ID: SWADESH_120
English word: to fly
Context: Past tense form: flew. The bird ___ over the building

EA01	1	ENG	طار	tˁAr	1
EA02	1	ENG	طار	tˁAr	1
GA01	1	ENG	طار	tˁAr	1
GA02	1	ENG	طار	tˁAr	1
LA01	1	ENG	طار	tˁAr	1
LA02	1	ENG	طار	tˁAr	1
MA01	1	ENG	طار	tˁAr	1
MA02	1	ENG	طار	tˁAr	1
SA01	1	ENG	طار	tˁAr	1

ID: SWADESH_121

English word: to walk
Context: Past tense form: walked. The boy ___ in the park

EA01	1	ENG	مِشي	miʃi	1
EA02	1	ENG	مِشي	miʃi	1
GA01	1	ENG	مِشا	miʃa	1
GA02	1	ENG	مَشا	maʃa	1
LA01	1	ENG	مَشا	maʃa	1
LA02	1	ENG	مِشي	miʃi	1
MA01	1	ENG	مشا	mʃa	1
MA02	1	ENG	مشا	mʃa	1
SA01	1	ENG	مَشا	maʃa	1

ID: SWADESH_122
English word: to come
Context: Past tense form: came. The boy ___ to the dinner

EA01	1	ENG	جِه	gih	1
EA02	1	ENG	جِه	gih	1
GA01	1	ENG	جا	ʤa	1
GA02	1	ENG	جا	ʤa	1
LA01	1	ENG	ءِجا	ʔiʤa	1
LA02	1	ENG	ءِجا	ʔiʒa	1
MA01	1	ENG	جا	ʒa	1
MA02	1	ENG	جا	ʒa	1
SA01	1	ENG	جاء	ʤAʔ	1

ID: SWADESH_123
English word: to lie
Context: Past tense form: lay. The boy ___ down on a bed

EA01	1	ENG	ءسْتْرَيَّح	ʔstəray+aħ	1
EA01	2	VAR	مَدِّد	mad+id	2
EA01	3	VAR	رَيَّح	ray+aħ	1
EA02	1	ENG	فَرَد	farad	3
EA02	2	ENG	رَيَّح	ray+aħ	1
EA02	3	VAR	مَدِّد	mad+id	2
EA02	4	VAR	ءسْتْرَيَّح	ʔstəray+aħ	1
GA01	1	ENG	ءِمبْطَح	ʔimbətˁaħ	4
GA01	2	VAR	ءِنسَدَح	ʔinsadaħ	5
GA01	3	VAR	تْمَدَّد	təmad+ad	2
GA02	1	ENG	نسَدَح	nsadaħ	5
GA02	2	ENG	مبَطَح	mbatˁaħ	4
LA01	1	ENG	تمَدَّد	tmad+ad	2
LA01	2	ENG	تبَطَّح	tbatˁ+aħ	4
LA01	3	VAR	تسَطَّح	tsatˁ+aħ	6

LA02	1	ENG	تسَطَّح	tsatˁ+aħ	6
MA01	1	ENG	تـجْبُّد	tʒəb+əd	7
MA01	2	VAR	تمْد	tməd	2
MA01	3	VAR	تكا	tka	8
MA02	1	ENG	تـجْبُّد	tʒəb+əd	7
MA02	2	VAR	تكا	tka	8
SA01	1	ENG	تَمَدَّد	tamad+ad	2
SA01	2	ENG	ءِستَراح	ʔistarAħ	1
SA01	3	ENG	ءِتَّكَاء	ʔit+akaʔ	8

ID: SWADESH_124
English word: to sit
Context: Past tense form: sat. The boy ___ on the chair

EA01	1	ENG	ءَعَد	ʔaʕad	1
EA02	1	ENG	ءَعَد	ʔaʕad	1
GA01	1	ENG	جْلَس	ʤəlas	2
GA01	2	ENG	گُعَد	gəʕad	1
GA02	1	ENG	گعَد	gaʕad	1
LA01	1	ENG	گعَد	gaʕad	1
LA02	1	ENG	ءَعَد	ʔaʕad	1
MA01	1	ENG	گلْس	gləs	2
MA02	1	ENG	گلْس	gləs	2
SA01	1	ENG	جَلَس	ʤalas	2
SA01	2	ENG	قَعَد	qaʕad	1

ID: SWADESH_125
English word: to stand
Context: Past tense form: stood. The boy ___ after he was sitting

EA01	1	ENG	وِءِف	wiʔif	1
EA01	2	VAR	ءام	ʔAm	2
EA02	1	ENG	وِءِف	wiʔif	1
EA02	2	VAR	ءام	ʔAm	2
GA01	1	ENG	وْگف	wəgaf	1
GA02	1	ENG	وَگف	wagaf	1
LA01	1	ENG	وْگِف	wəgif	1
LA01	2	VAR	گام	gAm	2
LA02	1	ENG	وِءْف	wiʔəf	1
LA02	2	VAR	ءام	ʔAm	2
MA01	1	ENG	وقَف	wqaf	1
MA02	1	ENG	وقَف	wqaf	1
SA01	1	ENG	وَقَف	waqaf	1
SA01	2	ENG	قام	qAm	2

ID: SWADESH_126
English word: to turn
Context: Past tense form: turned. The boy ___ around the building

EA01	1	ENG	لَف	laf	1
EA02	1	ENG	لَف	laf	1
GA01	1	ENG	لَف	laf	1
GA01	2	VAR	دار	dAr	2
GA01	3	VAR	ءِفتَر	ʔiftar	3
GA02	1	ENG	فتَر	ftar	3
LA01	1	ENG	لَف	laf	1
LA01	2	VAR	دار	dAr	2
LA02	1	ENG	بَرَم	baram	4
LA02	2	VAR	لَف	laf	1
LA02	3	VAR	دار	dAr	2
MA01	1	ENG	دار	dAr	2
MA02	1	ENG	دار	dAr	2
SA01	1	ENG	دار	dAr	2

ID: SWADESH_127
English word: to fall
Context: Past tense form: fell. A stone ___ from the top of a cliff

EA01	1	ENG	وِءِع	wiʔiʕ	1
EA02	1	ENG	وِءِع	wiʔiʕ	1
GA01	1	ENG	طاح	tˁAħ	2
GA02	1	ENG	طاح	tˁAħ	2
LA01	1	ENG	وِگّع	wigəʕ	1
LA02	1	ENG	وَءَع	waʔaʕ	1
LA02	2	ENG	وِعْع	wiʔəʕ	1
MA01	1	ENG	طاح	tˁAħ	2
MA02	1	ENG	طاح	tˁAħ	2
SA01	1	ENG	سَقَط	saqatˁ	3
SA01	2	ENG	وَقَع	waqaʕ	1

ID: SWADESH_128
English word: to give
Context: Past tense form: gave. The boy ___ the money to the man

EA01	1	ENG	ءِدّا	ʔid+a	1
EA02	1	ENG	ءِدّا	ʔid+a	1
GA01	1	ENG	عْطا	ʕətˁa	2
GA02	1	ENG	عَطا	ʕatˁa	2
LA01	1	ENG	ءَعطا	ʔaʕtˁa	2
LA02	1	ENG	عَطا	ʕatˁa	2
MA01	1	ENG	عطا	ʕtˁa	2

MA02	1	ENG	عطا	ʕtˤa	2
SA01	1	ENG	ءَعطا	ʔaʕtˤA	2

ID: SWADESH_129
English word: to hold
Context: Past tense form: held. The boy ___ the book in his hand

EA01	1	ENG	مِسِك	misik	1
EA02	1	ENG	مِسِك	misik	1
GA01	1	ENG	مْسَك	məsak	1
GA02	1	ENG	جَوَّد	ʤaw+ad	2
LA01	1	ENG	مَسَك	masak	1
LA02	1	ENG	هَدّا	had+a	3
LA02	2	ENG	مِسِك	misik	1
MA01	1	ENG	شْد	ʃəd	4
MA02	1	ENG	شْد	ʃəd	4
SA01	1	ENG	مَسَك	masak	1

ID: SWADESH_130
English word: to squeeze
Context: Past tense form: squeezed. The boy ___ the lemon

EA01	1	ENG	عَصَر	ʕasˤar	1
EA02	1	ENG	عَصَر	ʕasˤar	1
GA01	1	ENG	عْصَر	ʕəsˤar	1
GA02	1	ENG	عَصَر	ʕasˤar	1
LA01	1	ENG	عَصَر	ʕasˤar	1
LA02	1	ENG	عَصَر	ʕasˤar	1
MA01	1	ENG	عْصَّر	ʕəsˤ+ar	1
MA02	1	ENG	عْصَّر	ʕəsˤ+ar	1
SA01	1	ENG	عَصَر	ʕasˤar	1

ID: SWADESH_131
English word: to rub
Context: Past tense form: rubbed. The boy ___ the glass to clean it

EA01	1	ENG	دَعَك	daʕak	1
EA02	1	ENG	دَعَك	daʕak	1
GA01	1	ENG	مْسَح	məsaħ	2
GA02	1	ENG	فَرَك	farak	3
LA01	1	ENG	فَرَك	farak	3
LA01	2	VAR	دَعَك	daʕak	1
LA02	1	ENG	حَف	ħaf	4
MA01	1	ENG	مسَح	msaħ	2
MA02	1	ENG	مسَح	msaħ	2
SA01	1	ENG	فَرَك	farak	3

SA01	2	ENG	دَعَك	daʕak	1
SA01	3	ENG	مَسَح	masaħ	2

ID: SWADESH_132
English word: to wash
Context: Past tense form: washed. The boy ___ his hands

EA01	1	ENG	غَسَل	ʁasal	1
EA02	1	ENG	غَسَل	ʁasal	1
GA01	1	ENG	غْسَّل	ʁəs+al	1
GA02	1	ENG	غَسَل	ʁasal	1
LA01	1	ENG	غَسَّل	ʁas+al	1
LA01	2	ENG	غَسَل	ʁasal	1
LA02	1	ENG	غَسَّل	ʁas+al	1
MA01	1	ENG	غسْل	ʁsəl	1
MA02	1	ENG	غسْل	ʁsəl	1
SA01	1	ENG	غَسَل	ʁasal	1

ID: SWADESH_133
English word: to wipe
Context: Past tense form: wiped. The boy ___ the glasses with a cloth

EA01	1	ENG	مَسَح	masaħ	1
EA02	1	ENG	مَسَح	masaħ	1
GA01	1	ENG	مْسَح	məsaħ	1
GA02	1	ENG	مَش	maʃ	2
LA01	1	ENG	مَسَح	masaħ	1
LA02	1	ENG	مَسَح	masaħ	1
MA01	1	ENG	مسَح	msaħ	1
MA02	1	ENG	مسَح	msaħ	1
SA01	1	ENG	مَسَح	masaħ	1

ID: SWADESH_134
English word: to pull
Context: Past tense form: pulled. The boy ___ the toy car

EA01	1	ENG	شَد	ʃad	1
EA02	1	ENG	شَد	ʃad	1
EA02	2	VAR	جَر	gar	2
GA01	1	ENG	سَحَب	saħab	3
GA02	1	ENG	سَحَب	saħab	3
LA01	1	ENG	سَحَب	saħab	3
LA01	2	ENG	جَر	ʤar	2
LA02	1	ENG	شَد	ʃad	1
LA02	2	VAR	جَر	ʒar	2
MA01	1	ENG	جَر	ʒar	2

MA02	1	ENG	جَر	ʒar	2
SA01	1	ENG	شَدّ	ʃad+	1
SA01	2	ENG	جَرّ	ʤar+	2
SA01	3	ENG	سَحَب	saħab	3

ID: SWADESH_135
English word: to push
Context: Past tense form: pushed. The boy ___ the girl

EA01	1	ENG	زَء	zaʔ	1
EA02	1	ENG	زَء	zaʔ	1
GA01	1	ENG	دَف	daf	2
GA02	1	ENG	دَز	daz	3
LA01	1	ENG	دَفَش	dafaʃ	4
LA01	2	VAR	دَز	daz	3
LA02	1	ENG	دَفَش	dafaʃ	4
MA01	1	ENG	دفَع	dfaʕ	5
MA02	1	ENG	دفَع	dfaʕ	5
SA01	1	ENG	دَفَع	dafaʕ	5

ID: SWADESH_136
English word: to throw
Context: Past tense form: threw. The boy ___ a stone

EA01	1	ENG	رَما	rama	1
EA02	1	ENG	حَدَف	ħadaf	2
EA02	2	VAR	رَما	rama	1
GA01	1	ENG	حَذَف	ħaðaf	2
GA01	2	ENG	لاح	lAħ	3
GA01	3	ENG	نْطَل	nətˁal	4
GA02	1	ENG	رَما	rama	1
LA01	1	ENG	رَما	rama	1
LA01	2	VAR	زَت	zat	5
LA02	1	ENG	كَب	kab	6
LA02	2	ENG	زَت	zat	5
LA02	3	VAR	رَما	rama	1
MA01	1	ENG	رما	rma	1
MA01	2	VAR	لاح	lAħ	3
MA01	3	VAR	حذْف	ħdəf	2
MA02	1	ENG	لاح	lAħ	3
MA02	2	VAR	رما	rma	1
SA01	1	ENG	رَما	ramA	1
SA01	2	ENG	حَذَف	ħaðaf	2

ID: SWADESH_137

English word: to tie
Context: Past tense form: tied. The boy ___ the rope to the tree

EA01	1	ENG	رَبَط	rabatˤ	1
EA02	1	ENG	رَبَط	rabatˤ	1
GA01	1	ENG	رْبَط	rəbatˤ	1
GA02	1	ENG	رَبَط	rabatˤ	1
LA01	1	ENG	رَبَط	rabatˤ	1
LA01	2	VAR	عَگد	ʕagəd	2
LA02	1	ENG	بَكَّل	bak+al	3
LA02	2	VAR	رَبَط	rabatˤ	1
MA01	1	ENG	ربَط	rbatˤ	1
MA01	2	VAR	عقْد	ʕqəd	2
MA02	1	ENG	عقْد	ʕqəd	2
MA02	2	VAR	ربَط	rbatˤ	1
SA01	1	ENG	رَبَط	rabatˤ	1
SA01	2	ENG	عَقَد	ʕaqad	2

ID: SWADESH_138
English word: to sew
Context: Past tense form: sewed. The boy ___ the pieces of fabric

EA01	1	ENG	خَيَّط	xay+atˤ	1
EA02	1	ENG	خَيَّط	xay+atˤ	1
GA01	1	ENG	خَيَّط	xay+atˤ	1
GA02	1	ENG	خَيَّط	xay+atˤ	1
LA01	1	ENG	خَيَّط	xay+atˤ	1
LA02	1	ENG	خَيَّط	xay+atˤ	1
MA01	1	ENG	خيِّط	xy+itˤ	1
MA02	1	ENG	خيِّط	xy+itˤ	1
SA01	1	ENG	خَيَّط	xay+atˤ	1

ID: SWADESH_139
English word: to count
Context: Past tense form: counted. The boy ___ from one to ten

EA01	1	ENG	عَد	ʕad	1
EA02	1	ENG	عَد	ʕad	1
GA01	1	ENG	حَسب	ħasb	2
GA02	1	ENG	عَد	ʕad	1
LA01	1	ENG	عَد	ʕad	1
LA02	1	ENG	عَد	ʕad	1
MA01	1	ENG	حسْب	ħsəb	2
MA02	1	ENG	حسْب	ħsəb	2
SA01	1	ENG	عَدّ	ʕad+	1

ID: SWADESH_140
English word: to say
Context: Past tense form: said. The boy ___ that he will study

EA01	1	ENG	ءال	ʔAl	1
EA02	1	ENG	ءال	ʔAl	1
GA01	1	ENG	گال	gAl	1
GA02	1	ENG	گال	gAl	1
LA01	1	ENG	گال	gAl	1
LA02	1	ENG	ءال	ʔAl	1
MA01	1	ENG	گال	gAl	1
MA02	1	ENG	گال	gAl	1
MA02	2	VAR	قال	qAl	1
SA01	1	ENG	قال	qAl	1

ID: SWADESH_141
English word: to sing
Context: Past tense form: sang. The boy ___ at the concert

EA01	1	ENG	غَنّا	ʁan+a	1
EA02	1	ENG	غَنّا	ʁan+a	1
GA01	1	ENG	غَنّا	ʁan+a	1
GA02	1	ENG	غَنّا	ʁan+a	1
LA01	1	ENG	غَنّا	ʁan+a	1
LA02	1	ENG	غَنّا	ʁan+a	1
MA01	1	ENG	غَنّا	ʁan+a	1
MA02	1	ENG	غَنّا	ʁan+a	1
SA01	1	ENG	غنّا	ʁn+A	1

ID: SWADESH_142
English word: to play
Context: Past tense form: played. The boy ___ soccer with his friends

EA01	1	ENG	لِعِب	liʕib	1
EA02	1	ENG	لِعِب	liʕib	1
GA01	1	ENG	لْعَب	ləʕab	1
GA02	1	ENG	لَعَب	laʕab	1
LA01	1	ENG	لِعِب	liʕib	1
LA02	1	ENG	لِعِب	liʕib	1
MA01	1	ENG	لعْب	lʕəb	1
MA02	1	ENG	لعْب	lʕəb	1
SA01	1	ENG	لَعِب	laʕib	1

ID: SWADESH_143
English word: to float
Context: Past tense form: floated. The boat ___ on the water

EA01	1	ENG	عام	ʔAm	1
EA01	2	VAR	طَفا	tˁafa	2
EA02	1	ENG	عام	ʔAm	1
GA01	1	ENG	عام	ʔAm	1
GA02	1	ENG	طَفا	tˁafa	2
LA01	1	ENG	طَفا	tˁafa	2
LA01	2	VAR	عام	ʔAm	1
LA02	1	ENG	فاش	fAʃ	3
MA01	1	ENG	طفا	tˁfa	2
MA02	1	ENG	فلوطة	flUtˁa	4
SA01	1	ENG	عام	ʔAm	1
SA01	2	ENG	طَفا	tˁafA	2

ID: SWADESH_144
English word: to flow
Context: Past tense form: flowed. After it rained, the water ___ on the streets

EA01	1	ENG	جِري	giri	1
EA02	1	ENG	جِري	giri	1
GA01	1	ENG	جَرا	ʤara	1
GA01	2	VAR	مِشا	miʃa	2
GA02	1	ENG	غَرَّگ	ʁar+ag	3
LA01	1	ENG	سال	sAl	4
LA02	1	ENG	جَرا	ʒara	1
MA01	1	ENG	جرا	ʒra	1
MA02	1	ENG	جرا	ʒra	1
SA01	1	ENG	سال	sAl	4
SA01	2	ENG	جَرا	ʤarA	1

ID: SWADESH_145
English word: to freeze
Context: Past tense form: froze. The cold weather ___ the water

EA01	1	ENG	جَمِّد	gam+id	1
EA02	1	ENG	تَلِّج	tal+ig	2
EA02	2	VAR	جَمِّد	gam+id	1
GA01	1	ENG	جَمَّد	ʤam+ad	1
GA01	2	VAR	ثَلَّج	θal+adʒ	2
GA02	1	ENG	جَمَّد	ʤam+ad	1
GA02	2	VAR	ثَلَّج	θal+adʒ	2
LA01	1	ENG	جَمَّد	ʒam+ad	1
LA02	1	ENG	جَمَّد	ʒam+ad	1
LA02	2	ENG	جَلَّد	ʒal+ad	1
LA02	3	VAR	تَلَّج	tal+aʒ	2

MA01	1	ENG	جمَّد	ʒm+ad	1
MA02	1	ENG	جمَّد	ʒm+ad	1
MA02	2	VAR	تلْج	tl+əʒ	2
SA01	1	ENG	جَمَّد	ʤam+ad	1
SA01	2	ENG	جَلَّد	ʤal+ad	1

ID: SWADESH_146
English word: to swell
Context: Past tense form: swelled. My head ___ after I got hit

EA01	1	ENG	وِرِم	wirim	1
EA02	1	ENG	وِرِم	wirim	1
GA01	1	ENG	ءِنتْفَخ	ʔintəfax	2
GA02	1	ENG	ءِنتَفَخ	ʔintafax	2
LA01	1	ENG	وِرِم	wirim	1
LA02	1	ENG	وَرَّم	war+am	1
LA02	2	VAR	وِرِم	wirim	1
MA01	1	ENG	تنفَخ	tnfax	2
MA02	1	ENG	تنفَخ	tnfax	2
SA01	1	ENG	وَرِم	warim	1
SA01	2	ENG	ءِنتَفَخ	ʔintafax	2

ID: SWADESH_147
English word: sun
Context: This is the ___

EA01	1	ENG	شَمس	ʃams	1
EA02	1	ENG	شَمس	ʃams	1
GA01	1	ENG	شَمس	ʃams	1
GA02	1	ENG	شَمس	ʃams	1
LA01	1	ENG	شَمس	ʃams	1
LA02	1	ENG	شَمْس	ʃaməs	1
MA01	1	ENG	شَمش	ʃamʃ	1
MA02	1	ENG	شَمش	ʃamʃ	1
SA01	1	ENG	شَمس	ʃams	1

ID: SWADESH_148
English word: moon
Context: This is the ___

EA01	1	ENG	ءَمَر	ʔamar	1
EA02	1	ENG	ءَمَر	ʔamar	1
GA01	1	ENG	گُمَر	gəmar	1
GA02	1	ENG	قَمَر	qamar	1
LA01	1	ENG	گمَر	gamar	1
LA02	1	ENG	ءَمَر	ʔamar	1

MA01	1	ENG	قَمَر	qamar	1
MA02	1	ENG	ڭمرة	gəmra	1
SA01	1	ENG	قَمَر	qamar	1

ID: SWADESH_149
English word: star
Context: This is a ___ (in the sky)

EA01	1	ENG	نِجم	nigm	1
EA02	1	ENG	نِجم	nigm	1
EA02	2	ENG	نْجمة	nəgma	1
GA01	1	ENG	نَجِم	naʤim	1
GA02	1	ENG	نَجْم	naʤəm	1
LA01	1	ENG	نْجْم	nəʤəm	1
LA02	1	ENG	نِجمة	niʒmə	1
MA01	1	ENG	نْجمة	nəʒma	1
MA02	1	ENG	نْجمة	nəʒma	1
SA01	1	ENG	نَجم	naʤm	1

ID: SWADESH_150
English word: water
Context: This is a cup of ___

EA01	1	ENG	مَيّة	may+a	1
EA02	1	ENG	مَيّة	may+a	1
GA01	1	ENG	موية	mUya	1
GA02	1	ENG	ماي	mAy	1
LA01	1	ENG	مَي	may	1
LA02	1	ENG	مَي	may	1
MA01	1	ENG	ما	ma	1
MA02	1	ENG	ما	ma	1
SA01	1	ENG	ماء	mAʔ	1

ID: SWADESH_151
English word: rain
Context: This is ___

EA01	1	ENG	مَطَر	matˁar	1
EA02	1	ENG	مَطَر	matˁar	1
GA01	1	ENG	مْطَر	mətˁar	1
GA02	1	ENG	مَطَر	matˁar	1
LA01	1	ENG	شِتا	ʃita	2
LA02	1	ENG	شِتا	ʃitə	2
MA01	1	ENG	شتا	ʃta	2
MA02	1	ENG	شتا	ʃta	2
SA01	1	ENG	مَطَر	matˁar	1

ID: SWADESH_152
English word: river
Context: This is a ___

EA01	1	ENG	نَهر	nahr	1
EA02	1	ENG	نَهر	nahr	1
GA01	1	ENG	نْهَر	nəhar	1
GA02	1	ENG	نَهَر	nahar	1
LA01	1	ENG	نَهْر	nahər	1
LA02	1	ENG	نَهْر	nahər	1
MA01	1	ENG	واد	wAd	2
MA02	1	ENG	واد	wAd	2
SA01	1	ENG	نَهر	nahr	1

ID: SWADESH_153
English word: lake
Context: This is a ___

EA01	1	ENG	بُحَيرة	buħayra	1
EA02	1	ENG	بُحيرة	buħIra	1
GA01	1	ENG	بُحَيرة	buħayra	1
GA02	1	ENG	بُحَيرة	buħayra	1
LA01	1	ENG	بُحَيرة	buħayra	1
LA01	2	VAR	بْركة	bərkə	2
LA02	1	ENG	بُحَيرة	buħayra	1
MA01	1	ENG	بُحَيرة	buħayra	1
MA02	1	ENG	بْركة	bərka	2
MA02	2	ENG	بُحَيرة	buħayra	1
SA01	1	ENG	بُحَيرة	buħayra	1
SA01	2	ENG	بِركة	birka	2

ID: SWADESH_154
English word: sea
Context: This is a ___

EA01	1	ENG	بَحر	baħr	1
EA02	1	ENG	بَحر	baħr	1
GA01	1	ENG	بَحَر	baħar	1
GA02	1	ENG	بَحَر	baħar	1
LA01	1	ENG	بَحَر	baħar	1
LA02	1	ENG	بَحْر	baħər	1
MA01	1	ENG	بحْر	bħər	1
MA02	1	ENG	بحْر	bħər	1
SA01	1	ENG	بَحر	baħr	1

ID: SWADESH_155
English word: salt
Context: This is ___ (Mass noun)

EA01	1	ENG	مَلح	malħ	1
EA02	1	ENG	مَلح	malħ	1
GA01	1	ENG	مْلح	məlħ	1
GA02	1	ENG	مْلح	məlħ	1
LA01	1	ENG	مْلْح	mələħ	1
LA02	1	ENG	مْلْح	mələħ	1
MA01	1	ENG	مْلحة	məlħa	1
MA02	1	ENG	مْلحة	məlħa	1
SA01	1	ENG	مِلح	milħ	1

ID: SWADESH_156
English word: stone
Context: This is a ___

EA01	1	ENG	حَجَر	ħagar	1
EA02	1	ENG	حَجَر	ħagar	1
EA02	2	ENG	حِجارة	ħigAra	1
GA01	1	ENG	حَجَر	ħadʒar	1
GA01	2	VAR	حَصا	ħasˁa	2
GA02	1	ENG	حَجَر	ħadʒar	1
LA01	1	ENG	حَجَر	ħadʒar	1
LA02	1	ENG	حَجرَ	ħaʒra	1
MA01	1	ENG	حْجرة	ħəʒra	1
MA02	1	ENG	حْجرة	ħəʒra	1
SA01	1	ENG	حَجَر	ħadʒar	1
SA01	2	ENG	حَصا	ħasˁA	2

ID: SWADESH_157
English word: sand
Context: This is ___ (Mass noun)

EA01	1	ENG	رَمل	raml	1
EA02	1	ENG	رَمل	raml	1
GA01	1	ENG	رَمْل	ramәl	1
GA02	1	ENG	رَمْل	ramәl	1
LA01	1	ENG	رَمْل	ramәl	1
LA02	1	ENG	رَمْل	ramәl	1
MA01	1	ENG	رْملة	rәmla	1
MA02	1	ENG	رْملة	rәmla	1
SA01	1	ENG	رَمل	raml	1

ID: SWADESH_158

English word: dust
Context: This is ___ (on the furniture)

EA01	1	ENG	تُراب	turAb	1
EA02	1	ENG	تُراب	turAb	1
GA01	1	ENG	غبار	ʁbAr	2
GA02	1	ENG	غبار	ʁbAr	2
LA01	1	ENG	غَبرا	ʁabra	2
LA02	1	ENG	غَبرا	ʁabra	2
MA01	1	ENG	غْبرة	ʁəbra	2
MA02	1	ENG	غْبرة	ʁəbra	2
SA01	1	ENG	تُراب	turAb	1
SA01	2	ENG	غُبار	ʁubAr	2

ID: SWADESH_159
English word: earth
Context: This is good ___ for growing potatoes

EA01	1	ENG	عَرض	ʔardˁ	1
EA02	1	ENG	عَرض	ʔardˁ	1
GA01	1	ENG	عَرظ	ʔarðˁ	1
GA02	1	ENG	عَرظ	ʔarðˁ	1
LA01	1	ENG	عَرظ	ʔarðˁ	1
LA02	1	ENG	عَرْض	ʔarədˁ	1
MA01	1	ENG	عَرض	ʔardˁ	1
MA02	1	ENG	عَرض	ʔardˁ	1
SA01	1	ENG	عَرض	ʔardˁ	1

ID: SWADESH_160
English word: cloud
Context: This is a ___

EA01	1	ENG	سَحابة	saħAba	1
EA02	1	ENG	سَحابة	saħAba	1
GA01	1	ENG	غْيمة	ʁəymə	2
GA02	1	ENG	غْيمة	ʁəymə	2
LA01	1	ENG	غيمة	ʁImə	2
LA02	1	ENG	غَيمة	ʁaymə	2
MA01	1	ENG	سحابة	sħAba	1
MA01	2	VAR	غيمة	ʁIma	2
MA02	1	ENG	سحابة	sħAba	1
MA02	2	VAR	غيمة	ʁIma	2
SA01	1	ENG	سَحابة	saħAba	1
SA01	2	ENG	غَيمة	ʁayma	2

ID: SWADESH_161

English word: fog
Context: This is a ___

EA01	1	ENG	غَيام	ʁayAm	1
EA02	1	ENG	ضَباب	dˤabAb	2
GA01	1	ENG	ظَباب	ðˤabAb	2
GA02	1	ENG	ظَباب	ðˤabAb	2
LA01	1	ENG	ظَباب	ðˤabAb	2
LA02	1	ENG	ضَباب	dˤabAb	2
MA01	1	ENG	ضباب	dˤbAb	2
MA02	1	ENG	ضباب	dˤbAb	2
SA01	1	ENG	ضَباب	dˤabAb	2

ID: SWADESH_162
English word: sky
Context: The __ is blue

EA01	1	ENG	سَما	sama	1
EA02	1	ENG	سَما	sama	1
GA01	1	ENG	سْما	səma	1
GA02	1	ENG	سَما	sama	1
LA01	1	ENG	سَما	sama	1
LA02	1	ENG	سَما	sama	1
MA01	1	ENG	سما	sma	1
MA02	1	ENG	سما	sma	1
SA01	1	ENG	سَماء	samAʔ	1

ID: SWADESH_163
English word: wind
Context: The ___ is blowing

EA01	1	ENG	هَوا	hawa	1
EA02	1	ENG	هَوا	hawa	1
GA01	1	ENG	ريح	rIħ	2
GA01	2	VAR	هَوا	hawa	1
GA02	1	ENG	هَوا	hawa	1
LA01	1	ENG	ريح	rIħ	2
LA02	1	ENG	هَوا	hawa	1
MA01	1	ENG	ريح	rIħ	2
MA02	1	ENG	ريح	rIħ	2
MA02	2	VAR	هوا	hwa	1
SA01	1	ENG	ريح	rIħ	2
SA01	2	ENG	هَواء	hawAʔ	1

ID: SWADESH_164
English word: snow

Context: The ___ falling

EA01	1	ENG	تَلج	talg	1
EA02	1	ENG	تَلج	talg	1
GA01	1	ENG	ثَلج	θalʤ	1
GA02	1	ENG	ثَلج	θalʤ	1
LA01	1	ENG	تَلِج	taliʒ	1
LA02	1	ENG	تَلِج	taliʒ	1
MA01	1	ENG	ثْلج	təlʒ	1
MA02	1	ENG	ثْلج	təlʒ	1
SA01	1	ENG	ثَلج	θalʤ	1

ID: SWADESH_165
English word: ice
Context: This is a cube of ___

EA01	1	ENG	تَلج	talg	1
EA02	1	ENG	تَلج	talg	1
GA01	1	ENG	ثَلج	θalʤ	1
GA02	1	ENG	ثَلج	θalʤ	1
LA01	1	ENG	ثَلج	θalʤ	1
LA02	1	ENG	تَلِج	taliʒ	1
MA01	1	ENG	تلج	tlʒ	1
MA02	1	ENG	گلاصو	glAsˁu	2
SA01	1	ENG	ثَلج	θalʤ	1

ID: SWADESH_166
English word: smoke
Context: There may be a fire in that house, ___ is coming out of the windows

EA01	1	ENG	دُخّان	dux+An	1
EA02	1	ENG	دُخّان	dux+An	1
GA01	1	ENG	دْخّان	dəx+An	1
GA02	1	ENG	دْخّان	dəx+An	1
LA01	1	ENG	دُخّان	dux+An	1
LA02	1	ENG	دْخّان	dəx+An	1
MA01	1	ENG	دْخّان	dəx+An	1
MA02	1	ENG	دْخّان	dəx+An	1
SA01	1	ENG	دُخان	duxAn	1

ID: SWADESH_167
English word: fire
Context: This is a ___ (in the grill)

EA01	1	ENG	نار	nAr	1
EA02	1	ENG	نار	nAr	1

GA01	1	ENG	نار	nAr	1
GA02	1	ENG	نار	nAr	1
LA01	1	ENG	نار	nAr	1
LA02	1	ENG	نار	nAr	1
MA01	1	ENG	عافية	ʕAfya	2
MA02	1	ENG	عافية	ʕAfya	2
SA01	1	ENG	نار	nAr	1

ID: SWADESH_168
English word: ash
Context: This is ___ (in the grill)

EA01	1	ENG	رَماد	ramAd	1
EA02	1	ENG	رَماد	ramAd	1
GA01	1	ENG	رْماد	rəmAd	1
GA02	1	ENG	رماد	rmAd	1
LA01	1	ENG	سَچَن	sa¢an	2
LA02	1	ENG	صَفوة	sʕafwə	3
MA01	1	ENG	رماد	rmAd	1
MA02	1	ENG	رماد	rmAd	1
SA01	1	ENG	رَماد	ramAd	1

ID: SWADESH_169
English word: to burn
Context: Past tense form: burned. The grill ___ the boy's hand

EA01	1	ENG	حَرَء	ħaraʔ	1
EA02	1	ENG	حَرَء	ħaraʔ	1
GA01	1	ENG	حَرَگ	ħarag	1
GA02	1	ENG	حَرَگ	ħarag	1
LA01	1	ENG	حَرَگ	ħarag	1
LA02	1	ENG	حَرَء	ħaraʔ	1
MA01	1	ENG	حرْق	ħrəq	1
MA02	1	ENG	حرْگ	ħrəg	1
SA01	1	ENG	حَرَق	ħaraq	1

ID: SWADESH_170
English word: road
Context: This is a nice ___ between the two cities

EA01	1	ENG	طَريء	tʕarIʔ	1
EA02	1	ENG	طَريء	tʕarIʔ	1
GA01	1	ENG	طْريگ	tʕərIg	1
GA02	1	ENG	طَريگ	tʕarIg	1
LA01	1	ENG	طْريگ	tʕərIg	1
LA02	1	ENG	طَريء	tʕarIʔ	1

MA01	1	ENG	طريق	tˁrIq	1
MA02	1	ENG	طريق	tˁrIq	1
SA01	1	ENG	طَريق	tˁarIq	1

ID: SWADESH_171
English word: mountain
Context: This is a nice ___ it is very big and covered with snow

EA01	1	ENG	جَبَل	gabal	1
EA02	1	ENG	جَبَل	gabal	1
GA01	1	ENG	جْبَل	ʤəbal	1
GA02	1	ENG	جَبَل	ʤabal	1
LA01	1	ENG	جَبَل	ʤabal	1
LA02	1	ENG	جَبَل	ʒabal	1
MA01	1	ENG	جبْل	ʒbəl	1
MA02	1	ENG	جبْل	ʒbəl	1
SA01	1	ENG	جَبَل	ʤabal	1

ID: SWADESH_172
English word: red
Context: The color is ___

EA01	1	ENG	ءَحمَر	ʔaħmar	1
EA02	1	ENG	ءَحمَر	ʔaħmar	1
GA01	1	ENG	ءَحمَر	ʔaħmar	1
GA02	1	ENG	ءَحمَر	ʔaħmar	1
LA01	1	ENG	ءَحمَر	ʔaħmar	1
LA02	1	ENG	ءَحمَر	ʔaħmar	1
MA01	1	ENG	حمَر	ħmar	1
MA02	1	ENG	حمَر	ħmar	1
SA01	1	ENG	ءَحمَر	ʔaħmar	1

ID: SWADESH_173
English word: green
Context: The color is ___

EA01	1	ENG	ءَخضَر	ʔaxdˁar	1
EA02	1	ENG	ءَخضَر	ʔaxdˁar	1
GA01	1	ENG	ءَخظَر	ʔaxðˁar	1
GA02	1	ENG	ءَخظَر	ʔaxðˁar	1
LA01	1	ENG	ءَخظَر	ʔaxðˁar	1
LA02	1	ENG	ءَخضَر	ʔaxdˁar	1
MA01	1	ENG	خضَر	xdˁar	1
MA02	1	ENG	خضَر	xdˁar	1
SA01	1	ENG	ءَخضَر	ʔaxdˁar	1

ID: SWADESH_174
English word: yellow
Context: The color is ___

EA01	1	ENG	عَصفَر	ʔasˁfar	1
EA02	1	ENG	عَصفَر	ʔasˁfar	1
GA01	1	ENG	عَصفَر	ʔasˁfar	1
GA02	1	ENG	عَصفَر	ʔasˁfar	1
LA01	1	ENG	عَصفَر	ʔasˁfar	1
LA02	1	ENG	عَصفَر	ʔasˁfar	1
MA01	1	ENG	صفَر	sˁfar	1
MA02	1	ENG	صفَر	sˁfar	1
SA01	1	ENG	عَصفَر	ʔasˁfar	1

ID: SWADESH_175
English word: white
Context: The color is ___

EA01	1	ENG	عَبيَض	ʔabyadˁ	1
EA02	1	ENG	عَبيَض	ʔabyadˁ	1
GA01	1	ENG	عَبيَظ	ʔabyaðˁ	1
GA02	1	ENG	عَبيَظ	ʔabyaðˁ	1
LA01	1	ENG	عَبيَض	ʔabyadˁ	1
LA02	1	ENG	عَبيَض	ʔabyadˁ	1
MA01	1	ENG	بيْض	byədˁ	1
MA02	1	ENG	بيْض	byədˁ	1
SA01	1	ENG	عَبيَض	ʔabyadˁ	1

ID: SWADESH_176
English word: black
Context: The color is ___

EA01	1	ENG	عِسوِد	ʔiswid	1
EA02	1	ENG	عِسوِد	ʔiswid	1
GA01	1	ENG	عَسوَد	ʔaswad	1
GA02	1	ENG	عَسوَد	ʔaswad	1
LA01	1	ENG	عَسوَد	ʔaswad	1
LA02	1	ENG	عَسوَد	ʔaswad	1
MA01	1	ENG	كحْل	kħəl	2
MA02	1	ENG	كحْل	kħəl	2
SA01	1	ENG	عَسوَد	ʔaswad	1

ID: SWADESH_177
English word: night
Context: Indefinit singular form, During the Summer, the ___ is short

EA01	1	ENG	ليل	lll	1

EA02	1	ENG	ليل	lll	1
GA01	1	ENG	ليل	lll	1
GA02	1	ENG	ليل	lll	1
LA01	1	ENG	ليل	lll	1
LA02	1	ENG	ليل	lll	1
MA01	1	ENG	ليل	lll	1
MA02	1	ENG	ليل	lll	1
SA01	1	ENG	لَيل	layl	1

ID: SWADESH_178
English word: day
Context: Indefinit singular form, During the Winter, the ___ is short

EA01	1	ENG	يوم	yUm	1
EA01	2	VAR	نَهار	nahAr	2
EA02	1	ENG	نَهار	nahAr	2
EA02	2	VAR	يوم	yUm	1
GA01	1	ENG	نَهار	nahAr	2
GA02	1	ENG	يوم	yUm	1
LA01	1	ENG	نهار	nhAr	2
LA02	1	ENG	نهار	nhAr	2
MA01	1	ENG	نهار	nhAr	2
MA02	1	ENG	نهار	nhAr	2
SA01	1	ENG	نَهار	nahAr	2

ID: SWADESH_179
English word: year
Context: Indefinit singular form, The ___ 2012 was happy

EA01	1	ENG	سَنة	sana	1
EA02	1	ENG	سَنة	sana	1
GA01	1	ENG	سَنة	sana	1
GA02	1	ENG	سَنة	sana	1
LA01	1	ENG	سَنة	sanə	1
LA02	1	ENG	سِنة	sinə	1
MA01	1	ENG	عام	ʕAm	2
MA02	1	ENG	عام	ʕAm	2
SA01	1	ENG	سَنة	sana	1
SA01	2	ENG	عام	ʕAm	2

ID: SWADESH_180
English word: warm
Context: The food is not cold and it is not hot, it is ___

EA01	1	ENG	دافي	dAfi	1
EA02	1	ENG	دافي	dAfi	1

GA01	1	ENG	دافي	dAfi	1
GA02	1	ENG	دافي	dAfi	1
LA01	1	ENG	دافي	dAfi	1
LA02	1	ENG	دافي	dAfi	1
MA01	1	ENG	دافي	dAfi	1
MA02	1	ENG	دافي	dAfi	1
SA01	1	ENG	دافِء	dAfiʔ	1

ID: SWADESH_181
English word: cold
Context: The weather is ___ outside, wear a coat before going out

EA01	1	ENG	ساءِع	sAʔiʕ	1
EA01	2	VAR	بَرد	bard	2
EA02	1	ENG	بَرد	bard	2
EA02	2	VAR	ساءِع	sAʔiʕ	1
EA02	3	VAR	سَءعة	saʔʕa	1
GA01	1	ENG	بارِد	bArid	2
GA02	1	ENG	بارِد	bArid	2
LA01	1	ENG	بارِد	bArid	2
LA01	2	VAR	سَگعة	sagʕa	1
LA02	1	ENG	بارِد	bArid	2
LA02	2	VAR	صَءعة	sʿaʔʕə	1
MA01	1	ENG	بارْد	bArəd	2
MA02	1	ENG	بارْد	bArəd	2
SA01	1	ENG	بارِد	bArid	2

ID: SWADESH_182
English word: full
Context: This cup is ___

EA01	1	ENG	مَليان	malyAn	1
EA02	1	ENG	مَليان	malyAn	1
GA01	1	ENG	مَليان	malyAn	1
GA02	1	ENG	مَليان	malyAn	1
GA02	2	VAR	مَتروس	matrUs	2
LA01	1	ENG	مَليان	malyAn	1
LA02	1	ENG	مَليان	malyAn	1
MA01	1	ENG	عامْر	ʕAmər	3
MA02	1	ENG	عامْر	ʕAmər	3
SA01	1	ENG	مَلءان	malʔAn	1

ID: SWADESH_183
English word: new
Context: This is a ___

shoe

EA01	1	ENG	جِديد	gidId	1
EA02	1	ENG	جِديد	gidId	1
GA01	1	ENG	جُديد	ʤədId	1
GA02	1	ENG	جَديد	ʤadId	1
LA01	1	ENG	جديد	ʒdId	1
LA02	1	ENG	جديد	ʒdId	1
MA01	1	ENG	جديد	ʒdId	1
MA02	1	ENG	جديد	ʒdId	1
SA01	1	ENG	جَديد	ʤadId	1

ID: SWADESH_184
English word: old
Context: This is an ___ shoe

EA01	1	ENG	عَديم	ʔadIm	1
EA02	1	ENG	عَديم	ʔadIm	1
GA01	1	ENG	گُديم	gədIm	1
GA02	1	ENG	قَديم	qadIm	1
LA01	1	ENG	گديم	gadIm	1
LA02	1	ENG	عَديم	ʔadIm	1
MA01	1	ENG	قديم	qdIm	1
MA02	1	ENG	قديم	qdIm	1
SA01	1	ENG	قَديم	qadIm	1

ID: SWADESH_185
English word: good
Context: This is a ___ shoe, its comfortable

EA01	1	ENG	كوَيِّس	kway+is	1
EA02	1	ENG	حِلو	ħilu	2
EA02	2	VAR	كوَيِّس	kway+is	1
GA01	1	ENG	زين	zIn	3
GA02	1	ENG	زين	zIn	3
LA01	1	ENG	مليح	mlIħ	4
LA01	2	ENG	كوَيِّس	kway+is	1
LA02	1	ENG	منيح	mnIħ	4
MA01	1	ENG	مزيان	mzyAn	3
MA02	1	ENG	مزيان	mzyAn	3
SA01	1	ENG	جَيِّد	ʤay+id	5

ID: SWADESH_186
English word: bad
Context: This is a ___ shoe, its very uncomfortable

EA01	1	ENG	وِحِش	wiħiʃ	1

EA02	1	ENG	وِحِش	wiħiʃ	1
GA01	1	ENG	سْيِّء	səy+iʔ	2
GA02	1	ENG	موزين	mUzIn	3
LA01	1	ENG	عاطِل	ʕatˁil	4
LA02	1	ENG	سَيِّء	say+iʔ	2
MA01	1	ENG	خايب	xAyb	5
MA02	1	ENG	خايب	xAyb	5
SA01	1	ENG	سَيِّء	say+iʔ	2

ID: SWADESH_187
English word: rotten
Context: Don’t eat the bread or apple, Its ___ (its color has changed)

EA01	1	ENG	مِعَفِّن	miʕaf+in	1
EA01	2	VAR	بايِز،	bAyizˁ	2
EA02	1	ENG	بايِز،	bAyizˁ	2
EA02	2	VAR	مِعَفِّن	miʕaf+in	1
GA01	1	ENG	معَفِّن	mʕaf+in	1
GA02	1	ENG	عَفَن	ʕafan	1
LA01	1	ENG	معَفِّن	mʕaf+in	1
LA01	2	VAR	مخَمِّج	mxam+idʒ	3
LA02	1	ENG	معَفَّن	mʕaf+an	1
MA01	1	ENG	خسْر	xsər	4
MA01	2	VAR	خامِج	xAmiʒ	3
MA02	1	ENG	خايب	xAyb	5
MA02	2	ENG	خامِج	xAmiʒ	3
SA01	1	ENG	مُتَعَفِّن	mutaʕaf+in	1

ID: SWADESH_188
English word: dirty
Context: This chair is ___

EA01	1	ENG	وِسِخ	wisix	1
EA02	1	ENG	وِسِخ	wisix	1
GA01	1	ENG	وَصْخ	wasˁəx	1
GA02	1	ENG	وْصْخ	wəsˁəx	1
LA01	1	ENG	وْصْخ	wəsˁəx	1
LA02	1	ENG	موَسَّخ	mwas+ax	1
LA02	2	VAR	وِسِخ	wisix	1
MA01	1	ENG	موَسَّخ	mwas+ax	1
MA02	1	ENG	موَسّخ	mwas+ax	1
SA01	1	ENG	وَسِخ	wasix	1

ID: SWADESH_189
English word: straight

Context: draw a ___ line between the points

EA01	1	ENG	مْستَقيم	məstaqIm	1
EA02	1	ENG	مْستَقيم	məstaqIm	1
GA01	1	ENG	مْستَقيم	məstaqIm	1
GA02	1	ENG	سيدا	sIda	2
LA01	1	ENG	مُستَقيم	mustaqIm	1
LA02	1	ENG	جالِس	ʒAlis	3
MA01	1	ENG	مُستَقيم	mustaqIm	1
MA02	1	ENG	طويل	tˁwIl	4
SA01	1	ENG	مُستَقيم	mustaqIm	1

ID: SWADESH_190
English word: round
Context: Masculine form, This is a ___ dish

EA01	1	ENG	مْدَوَّر	mədaw+ar	1
EA02	1	ENG	مْدَوَّر	mədaw+ar	1
GA01	1	ENG	داءِري	dAʔiri	1
GA02	1	ENG	داءِري	dAʔiri	1
LA01	1	ENG	مدَوَّر	mdaw+ar	1
LA02	1	ENG	مدَوَّر	mdaw+ar	1
MA01	1	ENG	مْدَوَّر	mdəw+ar	1
MA01	2	VAR	داءِري	dAʔiri	1
MA02	1	ENG	مضْوَّر	mdˁəw+ar	1
SA01	1	ENG	داءِري	dAʔiri	1

ID: SWADESH_191
English word: sharp
Context: This is a ___ axe

EA01	1	ENG	حاد	ħAd	1
EA01	2	VAR	حامي	ħAmi	2
EA02	1	ENG	حامي	ħAmi	2
GA01	1	ENG	حاد	ħAd	1
GA02	1	ENG	حاد	ħAd	1
LA01	1	ENG	ماظي	mAðˁi	3
LA02	1	ENG	مرَوَّس	mraw+as	4
LA02	2	ENG	حَد	ħad	1
MA01	1	ENG	ماضي	mAdˁi	3
MA02	1	ENG	ماضي	mAdˁi	3
SA01	1	ENG	ماض	mAdˁ	3
SA01	2	ENG	حاد	ħAd	1

ID: SWADESH_192
English word: dull

Context: This is a ___ axe

EA01	1	ENG	تِلِم	tilim	1
EA02	1	ENG	بارِد	bArid	2
EA02	2	VAR	تِلِم	tilim	1
GA01	1	ENG	مهوبحاد	mhUbħAd	3
GA01	2	VAR	مُشحاد	muʃħAd	3
GA02	1	ENG	موحاد	mUħAd	3
LA01	1	ENG	مْشماظي	məʃmAðˁi	4
LA02	1	ENG	مَنّوحَد	man+Uħad	3
MA01	1	ENG	حافي	ħAfi	5
MA02	1	ENG	حافي	ħAfi	5
SA01	1	ENG	بارِد	bArid	2
SA01	2	ENG	ثالِم	θAlim	1

ID: SWADESH_193
English word: smooth
Context: This is a ___ surface

EA01	1	ENG	ناعِم	nAʕim	1
EA02	1	ENG	ناعِم	nAʕim	1
GA01	1	ENG	ءَملَس	ʔamlas	2
GA01	2	VAR	ناعِم	nAʕim	1
GA02	1	ENG	ناعِم	nAʕim	1
LA01	1	ENG	ناعِم	nAʕim	1
LA01	2	ENG	مْلِس	məlis	2
LA02	1	ENG	مالِس	mAlis	2
LA02	2	VAR	ناعِم	nAʕim	1
MA01	1	ENG	رطْب	rtˁəb	3
MA01	2	VAR	مْلْس	məlәs	2
MA02	1	ENG	رطْب	rtˁəb	3
SA01	1	ENG	ءَملَس	ʔamlas	2
SA01	2	ENG	مَلِس	malis	2
SA01	3	ENG	ناعِم	nAʕim	1

ID: SWADESH_194
English word: wet
Context: This is a ___ cloth, it is dripping water

EA01	1	ENG	مَبلول	mablUl	1
EA02	1	ENG	مَبلول	mablUl	1
GA01	1	ENG	مْبلول	məblUl	1
GA02	1	ENG	رْطِب	rətˁib	2
LA01	1	ENG	مَبلول	mablUl	1
LA02	1	ENG	مبَلَّل	mbal+al	1
MA01	1	ENG	فازگ	fAzg	3

MA02	1	ENG	فازگ	fAzg	3
SA01	1	ENG	بَلیل	balIl	1
SA01	2	ENG	مُبتَل	mubtal	1

ID: SWADESH_195
English word: dry
Context: This is a ___
cloth

EA01	1	ENG	ناشِف	nAʃif	1
EA02	1	ENG	ناشِف	nAʃif	1
GA01	1	ENG	ناشِف	nAʃif	1
GA02	1	ENG	ناشِف	nAʃif	1
GA02	2	ENG	جاف	ʤAf	2
LA01	1	ENG	ناشِف	nAʃif	1
LA02	1	ENG	ناشِف	nAʃif	1
MA01	1	ENG	ناشف	nAʃf	1
MA02	1	ENG	ناشف	nAʃf	1
SA01	1	ENG	جاف	ʤAf	2
SA01	2	ENG	ناشِف	nAʃif	1

ID: SWADESH_196
English word: correct
Context: Masculine form, This is a ___ answer

EA01	1	ENG	صَحیح	sˤaħIħ	1
EA01	2	VAR	مَزبوط	mazbUtˤ	2
EA01	3	VAR	صَح	sˤaħ	1
EA02	1	ENG	صَح	sˤaħ	1
EA02	2	VAR	صَحیح	sˤaħIħ	1
EA02	3	VAR	مَزبوط	mazbUtˤ	2
GA01	1	ENG	عْدِل	ʕədil	3
GA01	2	ENG	صَحیح	sˤaħIħ	1
GA02	1	ENG	صَحیح	sˤaħIħ	1
LA01	1	ENG	صَحیح	sˤaħIħ	1
LA01	2	VAR	مَزبوط	mazbUtˤ	2
LA02	1	ENG	صَح	sˤaħ	1
LA02	2	VAR	صَحیح	sˤaħIħ	1
LA02	3	VAR	مَزبوط	mazbUtˤ	2
MA01	1	ENG	صحیح	sˤħIħ	1
MA02	1	ENG	صحِح	sˤħiħ	1
SA01	1	ENG	صَحیح	sˤaħIħ	1

ID: SWADESH_197
English word: near

Context: The river is ___ my house

EA01	1	ENG	عُرَيِّب	ʔuray+ib	1
EA02	1	ENG	عُرَيِّب	ʔuray+ib	1
GA01	1	ENG	گريب	grIb	1
GA02	1	ENG	گريب	garIb	1
LA01	1	ENG	گريب	garIb	1
LA02	1	ENG	عَريب	ʔarIb	1
MA01	1	ENG	قريب	qrIb	1
MA02	1	ENG	قريب	qrIb	1
SA01	1	ENG	قَريب	qarIb	1

ID: SWADESH_198
English word:
far
Context: The river is ___ from my house

EA01	1	ENG	بِعيد	biʕId	1
EA02	1	ENG	بِعيد	biʕId	1
GA01	1	ENG	بعيد	bʕId	1
GA02	1	ENG	بَعيد	baʕId	1
LA01	1	ENG	بعيد	bʕId	1
LA02	1	ENG	بعيد	bʕId	1
MA01	1	ENG	بعيد	bʕId	1
MA02	1	ENG	بعيد	bʕId	1
SA01	1	ENG	بَعيد	baʕId	1

ID: SWADESH_199
English word: right
Context: I am to the ___ of the boy (Looking at a picture of boys)

EA01	1	ENG	يُمين	yəmIn	1
EA02	1	ENG	يِمين	yimIn	1
GA01	1	ENG	يُمين	yəmIn	1
GA02	1	ENG	يَمين	yamIn	1
LA01	1	ENG	يَمين	yamIn	1
LA02	1	ENG	يَمين	yamIn	1
MA01	1	ENG	ليمْن	lImən	1
MA02	1	ENG	ليمْن	lImən	1
SA01	1	ENG	يَمين	yamIn	1

ID: SWADESH_200
English word: left
Context: I am to the ___ of the boy (Looking at a picture of boys)

EA01	1	ENG	شِمال	ʃimAl	1
EA02	1	ENG	شِمال	ʃimAl	1

GA01	1	ENG	يِسار	yisAr	2
GA02	1	ENG	يَسار	yasAr	2
LA01	1	ENG	يَسار	yasAr	2
LA01	2	ENG	شمال	ʃmAl	1
LA02	1	ENG	شمال	ʃmAl	1
MA01	1	ENG	ليسْر	lIsər	2
MA02	1	ENG	ليسْر	lIsər	2
SA01	1	ENG	يَسار	yasAr	2
SA01	2	ENG	شِمال	ʃimAl	1

ID: SWADESH_201
English word:
at
Context: I will see you ___ the meter where I always park my car

EA01	1	ENG	عَند	ʕand	1
EA02	1	ENG	عَند	ʕand	1
GA01	1	ENG	عِند	ʕind	1
GA02	1	ENG	عِند	ʕind	1
LA01	1	ENG	عِند	ʕind	1
LA02	1	ENG	عَلا	ʕala	2
MA01	1	ENG	عِند	ʕind	1
MA02	1	ENG	عِند	ʕind	1
SA01	1	ENG	عِند	ʕind	1

ID: SWADESH_202
English word:
in
Context: He is ___ the city

EA01	1	ENG	في	fi	1
EA02	1	ENG	في	fi	1
GA01	1	ENG	فِ	fi	1
GA02	1	ENG	فِ	fi	1
LA01	1	ENG	بِ	bi	2
LA02	1	ENG	بِ	bi	2
MA01	1	ENG	فْ	fə	1
MA02	1	ENG	فْ	fə	1
SA01	1	ENG	في	fI	1
SA01	2	ENG	بِ	bi	2

ID: SWADESH_203
English word: with
Context: I am ___ my friends

EA01	1	ENG	مَع	maʕ	1

EA02	1	ENG	مَع	maʕ	1
GA01	1	ENG	مَعَ	maʕa	1
GA02	1	ENG	معَ	mʕa	1
LA01	1	ENG	مَع	maʕ	1
LA02	1	ENG	مَع	maʕ	1
MA01	1	ENG	معَ	mʕa	1
MA02	1	ENG	معَ	mʕa	1
SA01	1	ENG	مَعَ	maʕa	1

ID: SWADESH_204
English word: and
Context: Ali ___ Saleh are friends

EA01	1	ENG	ءو	ʔu	1
EA02	1	ENG	ءو	ʔu	1
GA01	1	ENG	ءو	ʔu	1
GA02	1	ENG	ءو	ʔu	1
LA01	1	ENG	ءو	ʔu	1
LA02	1	ENG	ءو	ʔu	1
MA01	1	ENG	ءو	ʔu	1
MA02	1	ENG	ءو	ʔu	1
SA01	1	ENG	وَ	wa	1

ID: SWADESH_205
English word: if
Context: Let me know ___ you can read this

EA01	1	ENG	لَو	law	1
EA01	2	VAR	ءِزا	ʔiza	2
EA02	1	ENG	لَو	law	1
EA02	2	VAR	ءِزا	ʔiza	2
GA01	1	ENG	ءِذا	ʔiða	2
GA01	2	VAR	لَو	law	1
GA02	1	ENG	ءِذا	ʔiða	2
LA01	1	ENG	لَو	law	1
LA01	2	ENG	ءِذا	ʔiða	2
LA02	1	ENG	ءِزا	ʔiza	2
MA01	1	ENG	ءِلا	ʔila	3
MA02	1	ENG	ءِلا	ʔila	3
SA01	1	ENG	ءِذا	ʔiðA	2
SA01	2	ENG	لَو	law	1

ID: SWADESH_206
English word: because

Context: I went to the clinic ___ I was sick

EA01	1	ENG	عَشان	ʕaʃAn	1
EA01	2	VAR	عَلَشان	ʕalaʃAn	1
EA02	1	ENG	عَشان	ʕaʃAn	1
EA02	2	VAR	عَلَشان	ʕalaʃAn	1
GA01	1	ENG	لْعَنَّ	ləʔan+a	2
GA01	2	VAR	عَشان	ʕaʃAn	1
GA02	1	ENG	لِعَنَّ	liʔan+a	2
GA02	2	VAR	عَشان	ʕaʃAn	1
LA01	1	ENG	عَشان	ʕaʃAn	1
LA01	2	ENG	لَعِن	laʔin	2
LA02	1	ENG	لَعَن	laʔan	2
MA01	1	ENG	لْحقاش	ləħqAʃ	3
MA01	2	VAR	حيت	ħIt	4
MA02	1	ENG	حيت	ħIt	4
SA01	1	ENG	لِعَنَّ	liʔan+a	2

ID: SWADESH_207
English word: name
Context: The ___ of the boy is Ali

EA01	1	ENG	ءِسم	ʔism	1
EA02	1	ENG	ءِسم	ʔism	1
GA01	1	ENG	ءِسْم	ʔisəm	1
GA02	1	ENG	ءِسْم	ʔisəm	1
LA01	1	ENG	ءِسْم	ʔisəm	1
LA02	1	ENG	ءِسْم	ʔisəm	1
MA01	1	ENG	سميت	smIt	1
MA02	1	ENG	سميت	smIt	1
SA01	1	ENG	ءِسم	ʔism	1

APPENDIX B

ENCODING THE MATHEMATICAL REPRESENTATION OF SOUND

Dimension 1: Place of articulation

Place of articulation	Abbreviation	Default value
bilabial	b	1
labiodental	l	0.9
dental	d	0.8
alveolar	a	0.7
postalveolar	e	0.6
palatal	p	0.5
central_vowel	c	0.4
velar	v	0.3
uvular	u	0.2
pharyngeal	r	0.1
glottal	g	0

Dimension 2: Degree of constriction at the place of articulation

Degree of constriction	Abbreviation	Default value
stop	s	0
fricative	f	0.2
approximant	t	0.4
high vowel	h	0.6
mid vowel	m	0.8
low vowel	w	1

Dimension 3: Voicing

Dimension 4: Nasal

Dimension 5: Lateral

Dimension 6: Trill/Flap

Dimension 7: Affricated

Dimension 8: Rounded

Dimension 9: Long_vowel. Long vowels expressed by upper case letters

Dimension 10: Geminated. Expressed by +

Dimension 11: Emphatic. Expressed by ˠ

The list of phonemes

Phone	Place of articulation	Degree of constriction	Voicing	Nasal	Lateral	Trill/Flap	Affricated	Rounded	Long_vowel	Geminated	Emphatic
b	b	s	1	0	0	0	0	0	0	0	0
t	a	s	0	0	0	0	0	0	0	0	0
¢	a	s	0	0	0	0	1	0	0	0	0
d	a	s	1	0	0	0	0	0	0	0	0
ʤ	a	s	1	0	0	0	1	0	0	0	0
k	v	s	0	0	0	0	0	0	0	0	0
g	v	s	1	0	0	0	0	0	0	0	0
q	u	s	0	0	0	0	0	0	0	0	0
ʔ	g	s	0	0	0	0	0	0	0	0	0
m	b	s	1	1	0	0	0	0	0	0	0
n	a	s	1	1	0	0	0	0	0	0	0
r	a	s	1	0	0	1	0	0	0	0	0
f	l	f	0	0	0	0	0	0	0	0	0
θ	d	f	0	0	0	0	0	0	0	0	0
ð	d	f	1	0	0	0	0	0	0	0	0
s	a	f	0	0	0	0	0	0	0	0	0
z	a	f	1	0	0	0	0	0	0	0	0
ʃ	e	f	0	0	0	0	0	0	0	0	0
ʒ	e	f	1	0	0	0	0	0	0	0	0
x	v	f	0	0	0	0	0	0	0	0	0
ʁ	u	f	1	0	0	0	0	0	0	0	0
ħ	r	f	0	0	0	0	0	0	0	0	0
ʕ	r	f	1	0	0	0	0	0	0	0	0
h	g	f	0	0	0	0	0	0	0	0	0
y	p	t	1	0	0	0	0	0	0	0	0
l	a	t	1	0	1	0	0	0	0	0	0
w	v	t	1	0	0	0	0	1	0	0	0
i	p	h	1	0	0	0	0	0	0	0	0
u	v	h	1	0	0	0	0	1	0	0	0
a	c	w	1	0	0	0	0	0	0	0	0
ə	c	m	1	0	0	0	0	0	0	0	0
I	p	h	1	0	0	0	0	0	1	0	0
U	v	h	1	0	0	0	0	1	1	0	0
A	c	w	1	0	0	0	0	0	1	0	0